

## 明 細 書

音声データを選択するための装置、方法およびプログラム

技術分野

本発明は、音声データ選択装置、音声データ選択方法及びプログラムに関する。

背景技術

音声を合成する手法として、録音編集方式と呼ばれる手法がある。録音編集方式は、駅の音声案内システムや、車載用のナビゲーション装置などに用いられている。

録音編集方式は、単語と、この単語を読み上げる音声を表す音声データとを対応付けておき、音声合成する対象の文章を単語に区切ってから、これらの単語に対応付けられた音声データを取得してつなぎ合わせる、という手法である。

この録音編集方式については、例えば、特開平 10-49193 号公報（以降、文献 1 と呼ぶ）において詳細に説明されている。

しかし、音声データを単につなぎ合わせた場合、音声データ同士の境界では通常、音声のピッチ成分の周波数が不連続的に変化する、等の理由で、合成音声が不自然なものとなる。

この問題を解決する手法としては、同一の音素を互いに異なった韻律で読み上げる音声を表す複数の音声データを用意し、一方で音声合成する対象の文章に韻律予測を施して、予測結果に合致する音声データを選び出してつなぎ合わせる、という手法が考えられる。

しかし、音声データを音素毎に用意して録音編集方式により自然

な合成音声を得ようとする、音声データを記憶する記憶装置には膨大な記憶容量が必要となり、小型軽量の装置を用いる必要がある用途には適さない。また、検索する対象のデータの量も膨大なものとなるから、高速な処理が要求される用途にも適さない。

また、韻律予測は極めて複雑な処理であるので、韻律予測を用いたこの手法を実現するには、処理能力が高いプロセッサなどを用い、あるいは長時間をかけて処理を行わせる必要がある。従ってこの手法は、構成が簡単な装置を用いた高速な処理が要求される用途には適さない。

この発明は、上記実状に鑑みてなされたものであり、簡単な構成で高速に自然な合成音声を得るための音声データ選択装置、音声データ選択方法及びプログラムを提供することを目的とする。

#### 発明の開示

(1) 上記発明目的を達成するために、本発明の音声データ選択装置は、第1の局面においては、基本的に、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出する検索手段と、索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する選択手段と、から構成される。

前記音声データ選択装置は、選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を更に備えていてもよい。

また、本発明の音声データ選択方法は、基本的に、音声の波形を表す音声データを複数記憶し、文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出し、および索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する、という一連の処理ステップを含む。

また、この発明のコンピュータプログラムは、コンピュータを、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出する検索手段と、索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する選択手段と、して機能させるためのものとなっている。

(2) 本発明の第2の局面においては、音声選択装置は、基本的に、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測する予測手段と、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する選択手段と、から構成される。

前記選択手段は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での1次回帰を行う回帰計算の結果に基づいて、当該音声データのピッチの時間変化と前記予測手段による予測の結果との相関の強さを特定するものであってもよい。

前記選択手段は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間の相関係数に基づいて、当該音声データのピッチの時間変化と前記予測手段による予測の結果との相関の強さを特定するものであってもよい。

また、この発明の別の音声選択装置は、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測する予測手段と、前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択する選択手段と、より構成されており、前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数より得られるようになっている。

前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での1次回帰により得られる1次関数の勾配からなって

いてもよい。

また、前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での1次回帰により得られる1次関数の切片からなっているいてもよい。

前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との間の相関係数からなっているいてもよい。

前記相関を表す数値は、音声データが表す音片のピッチの時間変化を表すデータを種々のビット数循環シフトしたものが表す関数と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果を表す関数との相関係数の最大値からなっているいてもよい。

前記記憶手段は、音声データの読みを表す表音データを、当該音声データに対応付けて記憶していてもよく、また前記選択手段は、前記文章内の音片の読みに合致する読みを表す表音データが対応付けられている音声データを、当該音片と読みが共通する音片の波形を表す音声データとして扱うものであってもよい。

前記音声選択装置は、選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を更に備えていてもよい。

前記音声選択装置は、前記文章内の音片のうち、前記選択手段が音声データを選択できなかった音片について、前記記憶手段が記憶する音声データを用いることなく、当該音片の波形を表す音声データを合成する欠落部分合成手段を備えていてもよく、前記音声合成手段は、前記選択手段が選択した音声データ及び前記欠落部分合成

手段が合成した音声データを互いに結合することにより、合成音声を表すデータを生成するものであってもよい。

また、この発明の音声選択方法は、音声の波形を表す音声データを複数記憶し、文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測し、および各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する、という一連の処理ステップを含む。

また、この発明の別の音声選択方法は、音声の波形を表す音声データを複数記憶し、文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測し、前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択するようになっており、前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数より得られるものである。

また、この発明のコンピュータプログラムは、コンピュータを、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測する予測手段と、各前記音声データのうちから、前記文章を構成する音片と読み

が共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する選択手段と、して機能させるためのものとなっている。

さらに、この発明の別のコンピュータプログラムは、コンピュータを、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測する予測手段と、前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択する選択手段として機能させるためのプログラムであって、前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数より得られるものである、となっている。

(3) 本発明の第3の局面においては、音声データ選択装置は、基本的に、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力する文章情報入力手段と、前記文章情報が表す文章内の音片と読みが共通する部分を有する音声データを索出する検索部と、前記索出されたそれぞれの音声データを文章情報が表す文章に従って接続した際に互いに隣接する音声データ同士の関係に基づいた所定の評価基準に従って評価値を求め、出力する音声データの組み合わせを当該評価値に基づいて選択する選択手段と、から構成される。

前記評価基準は、音声データが表す音声と韻律予測結果との相関及び互いに隣接する音声データ同士の関係を示す評価値を定める基準であって、前記評価値は、前記音声データが表す音声の特徴を示すパラメータ、前記音声データが表す音声を互いに結合して得られる音声の特徴を示すパラメータ、及び、発話時間長に関する特徴を示すパラメータのうち、少なくともいずれかを含む評価式に基づいて得られるものであるものであってもよい。

あるいは、前記評価基準は、音声データが表す音声と韻律予測結果との相関及び互いに隣接する音声データ同士の関係を示す評価値を定める基準であって、前記評価値は、前記音声データが表す音声を互いに結合して得られる音声の特徴を示すパラメータを含み、また、前記音声データが表す音声の特徴を示すパラメータと発話時間長に関する特徴を示すパラメータのうち、少なくともいずれかを含む評価式に基づいて得られるものであってもよい。

前記音声データが表す音声を互いに結合して得られる音声の特徴を示すパラメータは、前記文章情報が表す文章内の音片と読みが共通する部分を有する音声の波形を表す音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ選択した場合における、互いに隣接する音声データ同士の境界でのピッチの差に基づいて得られるものであってもよい。

前記音片データ選択装置は、文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測する予測手段を備えていてもよく、前記評価基準は、音声データが表す音声と前記韻律予測手段の韻律予測結果との相関ないし差異を示す評価値を定める



基準であって、前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び／又は、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数に基づいて得られるものであってもよい。

前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での 1 次回帰により得られる 1 次関数の勾配及び／又は切片からなっているとしてもよい。

前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との間の相関係数からなっているとしてもよい。

あるいは、前記相関を表す数値は、音声データが表す音片のピッチの時間変化を表すデータを種々のビット数循環シフトしたものが表す関数と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果を表す関数との相関係数の最大値からなっているとしてもよい。

前記記憶手段は、音声データの読みを表す表音データを、当該音声データに対応付けて記憶していてもよく、前記選択手段は、前記文章内の音片の読みに合致する読みを表す表音データが対応付けられている音声データを、当該音片と読みが共通する音片の波形を表す音声データとして扱うものであってもよい。

前記音片データ選択装置は、選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を

更に備えていてもよい。

前記音片データ選択装置は、前記文章内の音片のうち、前記選択手段が音声データを選択できなかった音片について、前記記憶手段が記憶する音声データを用いることなく、当該音片の波形を表す音声データを合成する欠落部分合成手段を備えていてもよく、前記音声合成手段は、前記選択手段が選択した音声データ及び前記欠落部分合成手段が合成した音声データを互いに結合することにより、合成音声を表すデータを生成するものであってもよい。

また、この発明の音声データ選択方法は、音声の波形を表す音声データを複数記憶し、文章を表す文章情報を入力し、前記文章情報が表す文章内の音片と読みが共通する部分を有する音声データを索出し、前記索出されたそれぞれの音声データを文章情報が表す文章に従って接続した際に互いに隣接する音声データ同士の関係に基づいた所定の評価基準に従って評価値を求め、出力する音声データの組み合わせを当該評価値に基づいて選択する、一速の処理ステップを含む。

また、この発明のコンピュータプログラムは、コンピュータを、音声の波形を表す音声データを複数記憶する記憶手段と、文章を表す文章情報を入力する文章情報入力手段と、前記文章情報が表す文章内の音片と読みが共通する部分を有する音声データを索出する検索部と、前記索出されたそれぞれの音声データを文章情報が表す文章に従って接続した際に互いに隣接する音声データ同士の関係に基づいた所定の評価基準に従って評価値を求め、出力する音声データの組み合わせを当該評価値に基づいて選択する選択手段と、して機能させるためのものとなっている。

### 図面の簡単な説明

第 1 図は、この発明の各実施の形態に係る音声合成システムの構成を示すブロック図である。

第 2 図は、この発明の第 1 の実施の形態における音片データベースのデータ構造を模式的に示す図である。

第 3 図の (a) は、音片についてのピッチ成分の周波数の予測結果と、この音片と読みが合致する音片の波形を表す音片データのピッチ成分の周波数の時間変化とを 1 次回帰させる処理を説明するためのグラフであり、同図 (b) は、相関係数を求めるために用いる予測結果データ及びピッチ成分データの値の一例を示すグラフである。

第 4 図は、この発明の第 2 の実施の形態における音片データベースのデータ構造を模式的に示す図である。

第 5 図の (a) は、定型メッセージの読みを示す図であり、同図 (b) は、音片編集部に供給された音片データのリストであり、同図 (c) は、先行する音片の末尾におけるピッチ成分の周波数と後続の音片の先頭におけるピッチ成分の周波数との差の絶対値を示す図であり、同図 (d) は、音片編集部がどの音片データを選択するかを示す図である。

第 6 図は、この発明の各実施の形態に係る音声合成システムの機能を行うパーソナルコンピュータがフリーテキストデータを取得した場合の処理を示すフローチャートである。

第 7 図は、この発明の各実施の形態に係る音声合成システムの機能を行うパーソナルコンピュータが配信文字列データを取得した場合の処理を示すフローチャートである。

第 8 図は、この発明の第 1 の実施の形態に係る音声合成システムの機能を行うパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

第 9 図は、この発明の第 2 の実施の形態に係る音声合成システムの機能を行うパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

第 10 図は、この発明の第 3 の実施の形態に係る音声合成システムの機能を行うパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

#### 発明を実施するための最良の形態

以下、この発明の実施の形態を、音声合成システムを例とし、図面を参照して説明する。

##### (第 1 の実施の形態)

第 1 図は、この発明の第 1 の実施の形態に係る音声合成システムの構成を示す図である。図示するように、この音声合成システムは、本体ユニット M と、音片登録ユニット R とにより構成されている。

本体ユニット M は、言語処理部 1 と、一般単語辞書 2 と、ユーザ単語辞書 3 と、音響処理部 4 と、検索部 5 と、伸長部 6 と、波形データベース 7 と、音片編集部 8 と、検索部 9 と、音片データベース 10 と、話速変換部 11 とにより構成されている。

言語処理部 1、音響処理部 4、検索部 5、伸長部 6、音片編集部 8、検索部 9 及び話速変換部 11 は、いずれも、CPU (Central

Processing Unit) や D S P (Digital Signal Processor) 等のプロセッサや、このプロセッサが実行するためのプログラムを記憶するメモリなどより構成されており、それぞれ後述する処理を行う。

なお、言語処理部 1、音響処理部 4、検索部 5、伸長部 6、音片編集部 8、検索部 9 及び話速変換部 11 の一部又は全部の機能を単一のプロセッサが行うようにしてもよい。

一般単語辞書 2 は、P R O M (Programmable Read Only Memory) やハードディスク装置等の不揮発性メモリより構成されている。一般単語辞書 2 には、表意文字（例えば、漢字など）を含む単語等と、この単語等の読みを表す表音文字（例えば、カナや発音記号など）とが、この音声合成システムの製造者等によって、あらかじめ互いに対応付けて記憶されている。

ユーザ単語辞書 3 は、E E P R O M (Electrically Erasable/Programmable Read Only Memory) やハードディスク装置等のデータ書き換え可能な不揮発性メモリと、この不揮発性メモリへのデータの書き込みを制御する制御回路とにより構成されている。なお、プロセッサがこの制御回路の機能を行ってもよく、言語処理部 1、音響処理部 4、検索部 5、伸長部 6、音片編集部 8、検索部 9 及び話速変換部 11 の一部又は全部の機能を行うプロセッサがユーザ単語辞書 3 の制御回路の機能を行うようにしてもよい。

ユーザ単語辞書 3 は、表意文字を含む単語等と、この単語等の読みを表す表音文字とを、ユーザの操作に従って外部より取得し、互いに対応付けて記憶する。ユーザ単語辞書 3 には、一般単語辞書 2 に記憶されていない単語等とその読みを表す表音文字とが格納され

ていれば十分である。

波形データベース 7 は、PROM やハードディスク装置等の不揮発性メモリより構成されている。波形データベース 7 には、表音文字と、この表音文字が表す単位音声の波形を表す波形データをエントロピー符号化して得られる圧縮波形データとが、この音声合成システムの製造者等によって、あらかじめ互いに対応付けて記憶されている。単位音声は、規則合成方式の手法で用いられる程度の短い音声であり、具体的には、音素や、VCV (Vowel-Consonant-Vowel) 音節などの単位で区切られる音声である。なお、エントロピー符号化される前の波形データは、例えば、PCM (Pulse Code Modulation) 化されたデジタル形式のデータからなっていればよい。

音片データベース 10 は、PROM やハードディスク装置等の不揮発性メモリより構成されている。

音片データベース 10 には、例えば、第 2 図に示すデータ構造を有するデータが記憶されている。すなわち、図示するように、音片データベース 10 に格納されているデータは、ヘッダ部 HDR、インデックス部 IDX、ディレクトリ部 DIR 及びデータ部 DAT の 4 種に分かれている。

なお、音片データベース 10 へのデータの格納は、例えば、この音声合成システムの製造者によりあらかじめ行われ、及び／又は、音片登録ユニット R が後述する動作を行うことにより行われる。

ヘッダ部 HDR には、音片データベース 10 を識別するデータや、インデックス部 IDX、ディレクトリ部 DIR 及びデータ部 DAT のデータ量、データの形式、著作権等の帰属などを示すデータが格

納される。

データ部 D A T には、音片の波形を表す音片データをエントロピー符号化して得られる圧縮音片データが格納されている。

なお、音片とは、音声のうち音素 1 個以上を含む連続した 1 区間をいい、通常は単語 1 個分又は複数個分の区間からなる。

また、エントロピー符号化される前の音片データは、上述の圧縮波形データの生成のためエントロピー符号化される前の波形データと同じ形式のデータ（例えば、P C M されたデジタル形式のデータ）からなっていればよい。

ディレクトリ部 D I R には、個々の圧縮音声データについて、

(A) この圧縮音片データが表す音片の読みを示す表音文字を表すデータ（音片読みデータ）、

(B) この圧縮音片データが格納されている記憶位置の先頭のアドレスを表すデータ、

(C) この圧縮音片データのデータ長を表すデータ、

(D) この圧縮音片データが表す音片の発声スピード（再生した場合の時間長）を表すデータ（スピード初期値データ）、

(E) この音片のピッチ成分の周波数の時間変化を表すデータ（ピッチ成分データ）、

が、互いに対応付けられた形で格納されている。（なお、音片データベース 1 0 の記憶領域にはアドレスが付されているものとする。）

なお、第 2 図は、データ部 D A T に含まれるデータとして、読みが「サイタマ」である音片の波形を表す、データ量 1 4 1 0 h バイトの圧縮音片データが、アドレス 0 0 1 A 3 6 A 6 h を先頭とする

論理的位置に格納されている場合を例示している。（なお、本明細書及び図面において、末尾に“h”を付した数字は16進数を表す。）

また、ピッチ成分データは、例えば、図示するように、音片のピッチ成分の周波数をサンプリングして得られたサンプル $Y(i)$ （サンプルの総数を $n$ として、 $i$ は $n$ 以下の正の整数）を表すデータであるものとする。

なお、上述の(A)～(E)のデータの集合のうち少なくとも(A)のデータ（すなわち音片読みデータ）は、音片読みデータが表す表音文字に基づいて決められた順位に従ってソートされた状態で（例えば、表音文字がカナであれば、五十音順に従って、アドレス降順に並んだ状態で）、音片データベース10の記憶領域に格納されている。

インデックス部IDXには、ディレクトリ部DIRのデータのおおよその論理的位置を音片読みデータに基づいて特定するためのデータが格納されている。具体的には、例えば、音片読みデータがカナを表すものであるとして、カナ文字と、先頭1字がこのカナ文字であるような音片読みデータがどのような範囲のアドレスにあるかを示すデータとが、互いに対応付けて格納されている。

なお、一般単語辞書2、ユーザ単語辞書3、波形データベース7及び音片データベース10の一部又は全部の機能を単一の不揮発性メモリが行うようにしてもよい。

音片データベース10へのデータの格納は、第1図に示す音片登録ユニットRにより行われる。音片登録ユニットRは、図示するように、収録音片データセット記憶部12と、音片データベース作成部13と、圧縮部14とにより構成されている。なお、音片登録ユ



ニットRは音片データベース10とは着脱可能に接続されていてもよく、この場合は、音片データベース10に新たにデータを書き込むときを除いては、音片登録ユニットRを本体ユニットMから切り離した状態で本体ユニットMに後述の動作を行わせてよい。

収録音片データセット記憶部12は、ハードディスク装置等のデータ書き換え可能な不揮発性メモリより構成されている。

収録音片データセット記憶部12には、音片の読みを表す表音文字と、この音片を人が実際に発声したものを集音して得た波形を表す音片データとが、この音声合成システムの製造者等によって、あらかじめ互いに対応付けて記憶されている。なお、この音片データは、例えば、PCM化されたデジタル形式のデータからなっていればよい。

音片データベース作成部13及び圧縮部14は、CPU等のプロセッサや、このプロセッサが実行するためのプログラムを記憶するメモリなどより構成されており、このプログラムに従って後述する処理を行う。

なお、音片データベース作成部13及び圧縮部14の一部又は全部の機能を単一のプロセッサが行うようにしてもよく、また、言語処理部1、音響処理部4、検索部5、伸長部6、音片編集部8、検索部9及び話速変換部11の一部又は全部の機能を行うプロセッサが音片データベース作成部13や圧縮部14の機能を更に行ってもよい。また、音片データベース作成部13や圧縮部14の機能を行うプロセッサが、収録音片データセット記憶部12の制御回路の機能を兼ねてもよい。

音片データベース作成部13は、収録音片データセット記憶部1

2より、互いに対応付けられている表音文字及び音片データを読み出し、この音片データが表す音声のピッチ成分の周波数の時間変化と、発声スピードとを特定する。

発声スピードの特定は、例えば、この音片データのサンプル数を数えることにより特定すればよい。

一方、ピッチ成分の周波数の時間変化は、例えば、この音片データにケプストラム解析を施すことにより特定すればよい。具体的には、例えば、音片データが表す波形を時間軸上で多数の小部分へと区切り、得られたそれぞれの小部分の強度を、元の値の対数（対数の底は任意）に実質的に等しい値へと変換し、値が変換されたこの小部分のスペクトル（すなわち、ケプストラム）を、高速フーリエ変換の手法（あるいは、離散変数をフーリエ変換した結果を表すデータを生成する他の任意の手法）により求める。そして、このケプストラムの極大値を与える周波数のうちの最小値を、この小部分におけるピッチ成分の周波数として特定する。

なお、ピッチ成分の周波数の時間変化は、例えば、特開2003-108172号公報に開示された手法に従って音片データをピッチ波形データへと変換してから、このピッチ波形データに基づいて特定するようにすると良好な結果が期待できる。具体的には、音片データをフィルタリングしてピッチ信号を抽出し、抽出されたピッチ信号に基づいて、音片データが表す波形を単位ピッチ長の区間へと区切り、各区間について、ピッチ信号との相関関係に基づいて位相のずれを特定して各区間の位相を揃えることにより、音片データをピッチ波形信号へと変換すればよい。そして、得られたピッチ波形信号を音片データとして扱い、ケプストラム解析を行う等するこ

とにより、ピッチ成分の周波数の時間変化を特定すればよい。

一方、音片データベース作成部 1 3 は、収録音片データセット記憶部 1 2 より読み出した音片データを圧縮部 1 4 に供給する。

圧縮部 1 4 は、音片データベース作成部 1 3 より供給された音片データをエントロピー符号化して圧縮音片データを作成し、音片データベース作成部 1 3 に返送する。

音片データの発声スピード及びピッチ成分の周波数の時間変化を特定し、この音片データがエントロピー符号化され圧縮音片データとなって圧縮部 1 4 より返送されると、音片データベース作成部 1 3 は、この圧縮音片データを、データ部 D A T を構成するデータとして、音片データベース 1 0 の記憶領域に書き込む。

また、音片データベース作成部 1 3 は、書き込んだ圧縮音片データが表す音片の読みを示すものとして収録音片データセット記憶部 1 2 より読み出した表音文字を、音片読みデータとして音片データベース 1 0 の記憶領域に書き込む。

また、書き込んだ圧縮音片データの、音片データベース 1 0 の記憶領域内での先頭のアドレスを特定し、このアドレスを上述の (B) のデータとして音片データベース 1 0 の記憶領域に書き込む。

また、この圧縮音片データのデータ長を特定し、特定したデータ長を、(C) のデータとして音片データベース 1 0 の記憶領域に書き込む。

また、この圧縮音片データが表す音片の発声スピード及びピッチ成分の周波数の時間変化を特定した結果を示すデータを生成し、スピード初期値データ及びピッチ成分データとして音片データベース 1 0 の記憶領域に書き込む。

次に、この音声合成システムの動作を説明する。

まず、言語処理部 1 が、この音声合成システムに音声を作成させる対象としてユーザが用意した、表意文字を含む文章（フリーテキスト）を記述したフリーテキストデータを外部から取得したとして説明する。

なお、言語処理部 1 がフリーテキストデータを取得する手法は任意であり、例えば、図示しないインターフェース回路を介して外部の装置やネットワークから取得してもよいし、図示しない記録媒体ドライブ装置にセットされた記録媒体（例えば、フロッピー（登録商標）ディスクやCD-ROMなど）から、この記録媒体ドライブ装置を介して読み取ってもよい。また、言語処理部 1 の機能を行っているプロセッサが、自ら実行している他の処理で用いたテキストデータを、フリーテキストデータとして、言語処理部 1 の処理へと引き渡すようにしてもよい。

フリーテキストデータを取得すると、言語処理部 1 は、このフリーテキストに含まれるそれぞれの表意文字について、その読みを表す表音文字を、一般単語辞書 2 やユーザ単語辞書 3 を検索することにより特定する。そして、この表意文字を、特定した表音文字へと置換する。そして、言語処理部 1 は、フリーテキスト内の表意文字がすべて表音文字へと置換した結果得られる表音文字列を、音響処理部 4 へと供給する。

音響処理部 4 は、言語処理部 1 より表音文字列を供給されると、この表音文字列に含まれるそれぞれの表音文字について、当該表音文字が表す単位音声の波形を検索するよう、検索部 5 に指示する。

検索部 5 は、この指示に応答して波形データベース 7 を検索し、

表音文字列に含まれるそれぞれの表音文字が表す単位音声の波形を表す圧縮波形データを索出する。そして、索出された圧縮波形データを伸長部 6 へと供給する。

伸長部 6 は、検索部 5 より供給された圧縮波形データを、圧縮される前の波形データへと復元し、検索部 5 へと返送する。検索部 5 は、伸長部 6 より返送された波形データを、検索結果として音響処理部 4 へと供給する。

音響処理部 4 は、検索部 5 より供給された波形データを、言語処理部 1 より供給された表音文字列内での各表音文字の並びに従った順序で、音片編集部 8 へと供給する。

音片編集部 8 は、音響処理部 4 より波形データを供給されると、この波形データを、供給された順序で互いに結合し、合成音声を表すデータ（合成音声データ）として出力する。フリーテキストデータに基づいて合成されたこの合成音声は、規則合成方式の手法により合成された音声に相当する。

なお、音片編集部 8 が合成音声データを出力する手法は任意であり、例えば、図示しない D/A（Digital-to-Analog）変換器やスピーカを介して、この合成音声データが表す合成音声を再生するようにしてもよい。また、図示しないインターフェース回路を介して外部の装置やネットワークに送出してもよいし、図示しない記録媒体ドライブ装置にセットされた記録媒体へ、この記録媒体ドライブ装置を介して書き込んでもよい。また、音片編集部 8 の機能を行っているプロセッサが、自ら実行している他の処理へと、合成音声データを引き渡すようにしてもよい。

次に、音響処理部 4 が、外部より配信された、表音文字列を表す

データ（配信文字列データ）を取得したとする。（なお、音響処理部 4 が配信文字列データを取得する手法も任意であり、例えば、言語処理部 1 がフリーテキストデータを取得する手法と同様の手法で配信文字列データを取得すればよい。）

この場合、音響処理部 4 は、配信文字列データが表す表音文字列を、言語処理部 1 より供給された表音文字列と同様に扱う。この結果、配信文字列データが表す表音文字列に含まれる表音文字に対応する圧縮波形データが検索部 5 により索出され、圧縮される前の波形データが伸長部 6 により復元される。復元された各波形データは、音響処理部 4 を介して音片編集部 8 へと供給され、音片編集部 8 が、この波形データを、配信文字列データが表す表音文字列内での各表音文字の並びに従った順序で互いに結合し、合成音声データとして出力する。配信文字列データに基づいて合成されたこの合成音声データも、規則合成方式の手法により合成された音声を表す。

次に、音片編集部 8 が、定型メッセージデータ及び発声スピードデータを取得したとする。

なお、定型メッセージデータは、定型メッセージを表音文字列として表すデータであり、発声スピードデータは、定型メッセージデータが表す定型メッセージの発声スピードの指定値（この定型メッセージを発声する時間長の指定値）を示すデータである。

また、音片編集部 8 が定型メッセージデータや発声スピードデータを取得する手法は任意であり、例えば、言語処理部 1 がフリーテキストデータを取得する手法と同様の手法で定型メッセージデータや発声スピードデータを取得すればよい。

定型メッセージデータ及び発声スピードデータが音片編集部 8 に

供給されると、音片編集部 8 は、定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データをすべて索出するよう、検索部 9 に指示する。

検索部 9 は、音片編集部 8 の指示に応答して音片データベース 10 を検索し、該当する圧縮音片データと、該当する圧縮音片データに対応付けられている上述の音片読みデータ、スピード初期値データ及びピッチ成分データとを索出し、索出された圧縮波形データを伸長部 6 へと供給する。1 個の音片につき複数の圧縮音片データが該当する場合も、該当する圧縮音片データすべてが、音声合成に用いられるデータの候補として索出される。一方、圧縮音片データを索出できなかった音片があった場合、検索部 9 は、該当する音片を識別するデータ（以下、欠落部分識別データと呼ぶ）を生成する。

伸長部 6 は、検索部 9 より供給された圧縮音片データを、圧縮される前の音片データへと復元し、検索部 9 へと返送する。検索部 9 は、伸長部 6 より返送された音片データと、索出された音片読みデータ、スピード初期値データ及びピッチ成分データとを、検索結果として話速変換部 11 へと供給する。また、欠落部分識別データを生成した場合は、この欠落部分識別データも話速変換部 11 へと供給する。

一方、音片編集部 8 は、話速変換部 11 に対し、話速変換部 11 に供給された音片データを変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致するようにすることを指示する。

話速変換部 11 は、音片編集部 8 の指示に応答し、検索部 9 より供給された音片データを指示に合致するように変換して、音片編集

部 8 に供給する。具体的には、例えば、検索部 9 より供給された音片データの元の時間長を、索出されたスピード初期値データに基づいて特定した上、この音片データをリサンプリングして、この音片データのサンプル数を、音片編集部 8 の指示したスピードに合致する時間長にすればよい。

また、話速変換部 11 は、検索部 9 より供給された音片読みデータ、スピード初期値データ及びピッチ成分データも音片編集部 8 に供給し、欠落部分識別データを検索部 9 より供給された場合は、更にこの欠落部分識別データも音片編集部 8 に供給する。

なお、発声スピードデータが音片編集部 8 に供給されていない場合、音片編集部 8 は、話速変換部 11 に対し、話速変換部 11 に供給された音片データを変換せずに音片編集部 8 に供給するよう指示すればよく、話速変換部 11 は、この指示に応答し、検索部 9 より供給された音片データをそのまま音片編集部 8 に供給すればよい。

音片編集部 8 は、話速変換部 11 より音片データ、音片読みデータ、スピード初期値データ及びピッチ成分データを供給されると、供給された音片データのうちから、定型メッセージを構成する音片の波形に最もよく近似できる波形を表す音片データを、音片 1 個につき 1 個ずつ選択する。

具体的には、まず、音片編集部 8 は、定型メッセージデータが表す定型メッセージに、例えば「藤崎モデル」や「T o B I (Tone and Break Indices)」等の韻律予測の手法に基づいた解析を加えることにより、この定型メッセージ内の各音片のピッチ成分の周波数の時間変化を予測する。そして、音片毎に、ピッチ成分の周波数の時間変化の予測結果をサンプリングしたものを表すデジタル形式のデ



ータ（以下、予測結果データと呼ぶ）を生成する。

次に、音片編集部 8 は、定型メッセージ内のそれぞれの音片について、この音片のピッチ成分の周波数の時間変化の予測結果を表す予測結果データと、この音片と読みが合致する音片の波形を表す音片データのピッチ成分の周波数の時間変化を表すピッチ成分データとの相関を求める。

より具体的には、音片編集部 8 は、話速変換部 11 より供給された各々のピッチ成分データについて、例えば、数式 1 の右辺に示す値  $\alpha$  及び数式 2 の右辺に示す値  $\beta$  を求める。

$$\alpha = \frac{\sum_{i=1}^n [\{X(i) - m_x\} \cdot \{Y(i) - m_y\}]}{\sum_{i=1}^n \{X(i) - m_x\}^2}$$

$$\text{ただし、} m_x = \sum_{i=1}^n \frac{X(i)}{n}, \quad m_y = \sum_{i=1}^n \frac{Y(i)}{n}$$

$$\beta = m_y - (\alpha \cdot m_x)$$

第 3 図 (a) に示すように、ある音片についての予測結果データ（サンプルの総数は  $n$  個とする）の  $i$  番目のサンプルの値  $X(i)$ （ $i$  は整数）の 1 次関数として、この音片と読みが合致する音片の波形を表す音片データについてのピッチ成分データ（サンプルの総数は  $n$  個とする）の  $i$  番目のサンプル  $Y(i)$  の値を 1 次回帰させ

た場合、この 1 次関数の勾配は  $\alpha$ 、切片は  $\beta$  となる。(勾配  $\alpha$  の単位は例えば [ヘルツ/秒] であればよく、切片  $\beta$  の単位は例えば [ヘルツ] であればよい。)

なお、同一の読みの音片について、予測結果データとピッチ成分データとでサンプルの総数が互いに異なる場合は、両者のうち一方(または両方)を、1 次補間やラグランジェ補間あるいはその他任意の手法により補間した上でリサンプリングし、両者のサンプルの総数を揃えてから相関を求めるようにすればよい。

一方、音片編集部 8 は、話速変換部 11 より供給されたスピード初期値データと、音片編集部 8 に供給された定型メッセージデータ及び発声スピードデータとを用いて、数式 3 の右辺の値  $d_t$  を求める。この値  $d_t$  は、音片データが表す音片の発声スピードと、この音片と読みが合致する定型メッセージ内の音片の発声スピードとの時間差を表す係数である。

$$d_t = |(X_t - Y_t) / Y_t|$$

(ただし、 $Y_t$  は音片データが表す音片の発声スピード、 $X_t$  はこの音片と読みが合致する定型メッセージ内の音片の発声スピード)

そして、音片編集部 8 は、1 次回帰により得られた上述の  $\alpha$  及び  $\beta$  の値と、上述の係数  $d_t$  とに基づいて、定型メッセージ内の音片の読みと一致する音片を表す音片データのうち、数式 4 の右辺の値(評価値)  $cost_1$  が最大となるものを選択する。

$$cost_1 = 1 / (W_1 |1 - \alpha| + W_2 |\beta| + d_t)$$

(ただし、 $W_1$  及び  $W_2$  は所定の正の係数)

音片のピッチ成分の周波数の時間変化の予測結果と、この音片と読みが合致する音片の波形を表す音片データのピッチ成分の周波数

の時間変化とが互いに近いほど、勾配  $\alpha$  の値は 1 に近くなり、従って、値  $|1 - \alpha|$  は 0 に近くなる。そして、評価値  $\text{cost}_1$  は、音片のピッチの予測結果と音片データのピッチとの相関が高いほど大きな値となるようにするため、値  $|1 - \alpha|$  の 1 次関数の逆数の形をとっているので、評価値  $\text{cost}_1$  は、値  $|1 - \alpha|$  が 0 に近くなるほど大きな値となる。

一方、音声の抑揚は、音片のピッチ成分の周波数の時間変化により特徴付けられる。従って、勾配  $\alpha$  の値は、音声の抑揚の差異を敏感に反映する性質を有する。

このため、合成されるべき音声について抑揚の正確さが重視される場合（例えば、電子メール等のテキストを読み上げる音声を合成する場合等）は、上述の係数  $W_1$  の値をなるべく大きくすることが望ましい。

これに対し、音片のピッチ成分の基本周波数（ベースピッチ周波数）の予測結果と、この音片と読みが合致する音片の波形を表す音片データのベースピッチ周波数とが互いに近いほど、切片  $\beta$  の値は 0 に近くなる。従って、切片  $\beta$  の値は、音声のベースピッチ周波数の差異を敏感に反映する性質を有する。一方、評価値  $\text{cost}_1$  は、値  $|\beta|$  の 1 次関数の逆数とみることにもできる形をとっているので、評価値  $\text{cost}_1$  は、値  $|\beta|$  が 0 に近くなるほど大きな値となる。

一方、音声のベースピッチ周波数は、音声の話者の声質を支配する要因であり、話者の性別による差異も顕著である。

このため、合成されるべき音声についてベースピッチ周波数の正確さが重視される場合（例えば、合成音声の話者の性別や声質を明確にする必要がある場合など）は、上述の係数  $W_2$  の値をなるべく

大きくすることが望ましい。

動作の説明に戻ると、音片編集部 8 は、定型メッセージ内の音片の波形に近い波形を表す音片データを選択する一方で、話速変換部 11 より欠落部分識別データも供給されている場合には、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出して音響処理部 4 に供給し、この音片の波形を合成するよう指示する。

指示を受けた音響処理部 4 は、音片編集部 8 より供給された表音文字列を、配信文字列データが表す表音文字列と同様に扱う。この結果、この表音文字列に含まれる表音文字が示す音声の波形を表す圧縮波形データが検索部 5 により索出され、この圧縮波形データが伸長部 6 により元の波形データへと復元され、検索部 5 を介して音響処理部 4 へと供給される。音響処理部 4 は、この波形データを音片編集部 8 へと供給する。

音片編集部 8 は、音響処理部 4 より波形データを返送されると、この波形データと、話速変換部 11 より供給された音片データのうち音片編集部 8 が特定したものとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する。

なお、話速変換部 11 より供給されたデータに欠落部分識別データが含まれていない場合は、音響処理部 4 に波形の合成を指示することなく直ちに、音片編集部 8 が特定した音片データを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力すればよい。

以上説明した、この音声合成システムでは、音素より大きな単位

であり得る音片の波形を表す音片データが、韻律の予測結果に基づいて、録音編集方式により自然につながわせられ、定型メッセージを読み上げる音声合成される。音片データベース 10 の記憶容量は、音素毎に波形を記憶する場合に比べて小さくでき、また、高速に検索できる。このため、この音声合成システムは小型軽量に構成することができ、また高速な処理にも追従できる。

また、音片の波形の予測結果と音片データとの相関を複数の評価基準（例えば、1 次回帰させた場合の勾配や切片による評価と、音片の時間差による評価、など）で評価した場合は、これらの評価の結果に食い違いが生じる場合が多々あり得る。しかし、この音声合成システムでは、複数の評価基準で評価した結果が 1 個の評価値に基づいて総合され、適正な評価が行われる。

なお、この音声合成システムの構成は上述のものに限られない。

例えば、波形データや音片データは P C M 形式のデータである必要はなく、データ形式は任意である。

また、波形データベース 7 や音片データベース 10 は波形データや音片データを必ずしもデータ圧縮された状態で記憶している必要はない。波形データベース 7 や音片データベース 10 が波形データや音片データをデータ圧縮されていない状態で記憶している場合、本体ユニット M は伸長部 6 を備えている必要はない。

また、音片データベース作成部 13 は、図示しない記録媒体ドライブ装置にセットされた記録媒体から、この記録媒体ドライブ装置を介して、音片データベース 10 に追加する新たな圧縮音片データの材料となる音片データや表音文字列を読み取ってもよい。

また、音片登録ユニット R は、必ずしも収録音片データセット記

憶部 1 2 を備えている必要はない。

また、音片編集部 8 は、特定の音片の韻律を表す韻律登録データをあらかじめ記憶し、定型メッセージにこの特定の音片が含まれている場合は、この韻律登録データが表す韻律を、韻律予測の結果として扱うようにしてもよい。

また、音片編集部 8 は、過去の韻律予測の結果を韻律登録データとして新たに記憶するようにしてもよい。

また、音片編集部 8 は、上述の  $\alpha$  及び  $\beta$  の値を求める代わりに、話速変換部 1 1 より供給された各々のピッチ成分データについて、例えば、数式 5 の右辺に示す値  $R_{xy}(j)$  を、 $j$  の値を 0 以上  $n$  未満の各整数として、合計  $n$  個求め、得られた  $R_{xy}(0)$  から  $R_{xy}(n-1)$  までの  $n$  個の相関係数のうちの最大値を特定するようにしてもよい。

$$R_{xy}(j) = \frac{\sum_{i=1}^n [\{X(i) - m_x\} \cdot \{Y_j(i) - m_y\}]}{\sqrt{\sum_{i=1}^n \{X(i) - m_x\}^2} \sqrt{\sum_{i=1}^n \{Y(i) - m_y\}^2}}$$

$R_{xy}(j)$  は、ある音片についての予測結果データ（サンプル総数  $n$  個。なお、数式 5 における  $X(i)$  は数式 1 におけるものと同一である）と、この音片と読みが合致する音片の波形を表す音片データについてのピッチ成分データ（サンプルの総数  $n$  個）を一定の方向へ  $j$  個循環シフトして得られたサンプルの列（なお、数式 5 において  $Y_j(i)$  は、このサンプルの列の  $i$  番目のサンプルの値である）との相関係数の値である。

なお、第3図(b)は、 $R_{xy}(0)$  及び  $R_{xy}(j)$  の値を求めるために用いる予測結果データ及びピッチ成分データの値の一例を示すグラフである。ただし、 $Y(p)$  の値（ただし、 $p$  は1以上  $n$  以下の整数）は、循環シフトを行う前のピッチ成分データの  $p$  番目のサンプルの値である。従って、例えば、音片データのサンプルが時刻の早い順に並んでおり、循環シフトが下位方向（つまり時刻が遅い方）へと行われるものとすれば、 $j < p$  の場合は  $Y_j(p) = Y(p-j)$  であり、一方、 $1 \leq p \leq j$  の場合は  $Y_j(p) = Y(n-j+p)$  である。

そして、音片編集部8は、上述の  $R_{xy}(j)$  の最大値と、上述の係数  $d_t$  とに基づいて、定型メッセージ内の音片の読みと一致する音片を表す音片データのうち、数式6の右辺の値（評価値） $cost_2$  が最大となるものを選択すればよい。

$$cost_2 = 1 / (W_3 | R_{max} | + d_t)$$

（ただし、 $W_3$  は所定の係数、 $R_{max}$  は  $R_{xy}(0) \sim R_{xy}(n-1)$  のうちの最大値）

なお、音片編集部8は、必ずしもピッチ成分データを種々循環シフトしたものについて上述の相関係数を求める必要はなく、例えば、 $R_{xy}(0)$  の値をそのまま相関係数の最大値として扱うようにしてもよい。

また、評価値  $cost_1$  や  $cost_2$  は、係数  $d_t$  の項を含まなくてもよく、この場合、音片編集部8は、係数  $d_t$  を求める必要がない。

あるいは、音片編集部8は、係数  $d_t$  の値をそのまま評価値として用いてもよく、この場合、音片編集部は、勾配  $\alpha$  や、切片  $\beta$  や、

$R \times y(j)$  の値を求める必要がない。

また、ピッチ成分データは音片データが表す音片のピッチ長の時間変化を表すデータであってもよい。この場合、音片編集部 8 は、予測結果データとして、音片のピッチ長の時間変化の予測結果を表すデータを作成するものとし、この音片と読みが合致する音片の波形を表す音片データのピッチ長の時間変化を表すピッチ成分データとの相関を求めるようにすればよい。

また、音片データベース作成部 13 は、マイクロフォン、増幅器、サンプリング回路、A/D (Analog-to-Digital) コンバータ及び PCM エンコーダなどを備えていてもよい。この場合、音片データベース作成部 13 は、収録音片データセット記憶部 12 より音片データを取得する代わりに、自己のマイクロフォンが集音した音声を表す音声信号を増幅し、サンプリングして A/D 変換した後、サンプリングされた音声信号に PCM 変調を施すことにより、音片データを作成してもよい。

また、音片編集部 8 は、音響処理部 4 より返送された波形データを話速変換部 11 に供給することにより、当該波形データが表す波形の時間長を、発声スピードデータが示すスピードに合致させるようにしてもよい。

また、音片編集部 8 は、例えば、言語処理部 1 と共にフリーテキストデータを取得し、このフリーテキストデータが表すフリーテキストに含まれる音片の波形に最も近い波形を表す音片データを、定型メッセージに含まれる音片の波形に最も近い波形を表す音片データを選択する処理と実質的に同一の処理を行うことによって選択して、音声の合成に用いてもよい。



この場合、音響処理部 4 は、音片編集部 8 が選択した音片データが表す音片については、この音片の波形を表す波形データを検索部 5 に索出させなくてもよい。なお、音片編集部 8 は、音響処理部 4 が合成しなくてよい音片を音響処理部 4 に通知し、音響処理部 4 はこの通知に応答して、この音片を構成する単位音声の波形の検索を中止するようにすればよい。

また、音片編集部 8 は、例えば、音響処理部 4 と共に配信文字列データを取得し、この配信文字列データが表す配信文字列に含まれる音片の波形に最も近い波形を表す音片データを、定型メッセージに含まれる音片の波形に最も近い波形を表す音片データを選択する処理と実質的に同一の処理を行うことによって選択して、音声の合成に用いてもよい。この場合、音響処理部 4 は、音片編集部 8 が選択した音片データが表す音片については、この音片の波形を表す波形データを検索部 5 に索出させなくてもよい。

#### (第 2 の実施の形態)

次に、この発明の第 2 の実施の形態を説明する。この発明の第 2 の実施の形態に係る音声合成システムの物理的構成は、上述した第 1 の実施の形態における構成と実質的に同一である。

ただし、第 2 の実施の形態の音声合成システムにおける音片データベース 10 のディレクトリ部 D I R には、例えば第 4 図に示すように、個々の圧縮音声データについて、上述の (A) ~ (D) のデータが互いに対応づけられた形で格納されているほか、上述の (E) のデータに代え、ピッチ成分データとして、(F) この圧縮音片データが表す音片の先頭と末尾におけるピッチ成分の周波数を表すデータが、これら (A) ~ (D) のデータに対応付けられた形で格納

されている。

なお、第4図は、第2図と同様、データ部DATに含まれるデータとして、読みが「サイタマ」である音片の波形を表す、データ量1410hバイトの圧縮音片データが、アドレス001A36A6hを先頭とする論理的位置に格納されている場合を例示している。) また、上述の(A)～(D)及び(F)のデータの集合のうち少なくとも(A)のデータは、音片読みデータが表す表音文字に基づいて決められた順位に従ってソートされた状態で音片データベース10の記憶領域に格納されているものとする。

そして、音片登録ユニットRの音片データベース作成部13は、収録音片データセット記憶部12より、互いに対応付けられている表音文字及び音片データを読み出すと、この音片データが表す音声の発声スピードと、先頭及び末尾でのピッチ成分の周波数とを特定するものとする。

そして、読み出した音片データを圧縮部14に供給し、圧縮音片データの返送を受けると、この圧縮音片データ、収録音片データセット記憶部12より読み出した表音文字、この圧縮音片データの音片データベース10の記憶領域内での先頭のアドレス、この圧縮音片データのデータ長、及び、特定した発声スピードを示すスピード初期値データを、第1の実施の形態の音片データベース作成部13と同様の動作を行うことにより音片データベース10の記憶領域に書き込み、また、音声の先頭及び末尾におけるピッチ成分の周波数を特定した結果を示すデータを生成して、ピッチ成分データとして音片データベース10の記憶領域に書き込むものとする。

なお、発声スピード及びピッチ成分の周波数の特定は、例えば、

第 1 の実施の形態の音片データベース作成部 13 が行う手法と実質的に同一の手法により行えばよい。

次に、この音声合成システムの動作を説明する。

この音声合成システムの言語処理部 1 がフリーテキストデータを外部から取得した場合、及び、音響処理部 4 が配信文字列データを取得した場合の動作は、第 1 の実施の形態の音声合成システムが行う動作と実質的に同一である。（なお、言語処理部 1 がフリーテキストデータを取得する手法や音響処理部 4 が配信文字列データを取得する手法はいずれも任意であり、例えば、いずれも第 1 の実施の形態における言語処理部 1 や音響処理部 4 が行う手法と同様の手法によりフリーテキストデータあるいは配信文字列データを取得すればよい。）

次に、音片編集部 8 が、定型メッセージデータ及び発声スピードデータを取得したとする。（なお、音片編集部 8 が定型メッセージデータや発声スピードデータを取得する手法も任意であり、例えば、第 1 の実施の形態の音片編集部 8 が行う手法と同様の手法で定型メッセージデータや発声スピードデータを取得すればよい。）

定型メッセージデータ及び発声スピードデータが音片編集部 8 に供給されると、音片編集部 8 は、第 1 の実施の形態における音片編集部 8 と同様に、定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データをすべて索出するよう、検索部 9 に指示する。また、話速変換部 11 に対しても、第 1 の実施の形態における音片編集部 8 と同様に、話速変換部 11 に供給される音片データを変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致す

るようにすることを指示する。

すると、検索部 9、伸張部 6 及び話速変換部 11 が、第 1 の実施の形態における検索部 9、伸張部 6 及び話速変換部 11 の動作と実質的に同一の動作を行い、この結果、話速変換部 11 から音片編集部 8 へと、音片データ、音片読みデータ及びピッチ成分データが供給される。また、欠落部分識別データが検索部 9 より話速変換部 11 へと供給された場合は、更にこの欠落部分識別データも音片編集部 8 へと供給される。

音片編集部 8 は、話速変換部 11 より音片データ、音片読みデータ及びピッチ成分データを供給されると、以下説明する手順に従い、供給された音片データのうちから、定型メッセージを構成する音片の波形とみなせる波形を表す音片データを、音片 1 個につき 1 個ずつ選択する。

具体的には、まず、音片編集部 8 は、話速変換部 11 より供給されたピッチ成分データに基づき、話速変換部 11 より供給された各音片データの先頭及び末尾の各時点でのピッチ成分の周波数を特定する。そして、話速変換部 11 より供給された音片データのうちから、定型メッセージ内で隣接する音片同士の境界でのピッチ成分の周波数の差の絶対値を定型メッセージ全体で累計した値が最小になる、という条件を満たすように、音片データを選択する。

音片データを選択する条件を、第 5 図 (a) ~ (d) を参照して説明する。例えば、第 5 図 (a) に示すような、「このさきみぎかーぶです」という読みの定型メッセージを表す定型メッセージデータが音片編集部 8 に供給されたものとし、この定型メッセージが「このさき」、「みぎかーぶ」及び「です」という 3 個の音片からなる

ものとする。そして、第5図(b)にリストを示すように、音片データベース10が、読みが「このさき」である圧縮音片データが3個(第5図(b)において「A1」「A2」あるいは「A3」として表したもの)、読みが「みぎかーぶ」である圧縮音片データが2個(第5図(b)において「B1」あるいは「B2」として表したもの)、読みが「です」である圧縮音片データが3個(第5図(b)において「C1」「C2」あるいは「C3」として表したもの)、それぞれ索出され、伸長され、音片データとして音片編集部8へと供給されたとする。

一方、読みが「このさき」である各音片データが表す各音片の末尾におけるピッチ成分の周波数と読みが「みぎかーぶ」である各音片データが表す各音片の先頭におけるピッチ成分の周波数との差の絶対値は第5図(c)に示す通りであったとする。(第5図(c)は、例えば、音片データA1が表す音片の末尾におけるピッチ成分の周波数と音片データB1が表す音片の先頭におけるピッチ成分の周波数との差の絶対値は「123」であることを示している。なお、この絶対値の単位は、例えば「ヘルツ」である。)

また、読みが「みぎかーぶ」である各音片データが表す各音片の末尾におけるピッチ成分の周波数と読みが「です」である各音片データが表す各音片の先頭におけるピッチ成分の周波数との差の絶対値は第5図(c)に示す通りであったとする。

この場合において、「このさきみぎかーぶです」という定型メッセージを読み上げる音声の波形を音片データを用いて生成した場合、隣接する音片同士の境界でのピッチ成分の周波数の差の絶対値の累計が最小になる組み合わせは、A3、B2及びC2という組み合わせ

せである。従ってこの場合、音片編集部 8 は、第 5 図 (d) に示すように、音片データ A 3、B 2 及び C 2 を選択する。

この条件を満たす音片データを選択するために、音片編集部 8 は、例えば、定型メッセージ内で隣接する音片同士の境界でのピッチ成分の周波数の差の絶対値を距離として定義し、D P (Dynamic Programming) マッチングの手法により音片データを選ぶようにすればよい。

一方、音片編集部 8 は、話速変換部 11 より欠落部分識別データも供給されている場合には、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出して音響処理部 4 に供給し、この音片の波形を合成するよう指示する。

指示を受けた音響処理部 4 は、音片編集部 8 より供給された表音文字列を、配信文字列データが表す表音文字列と同様に扱う。この結果、この表音文字列に含まれる表音文字が示す音声の波形を表す圧縮波形データが検索部 5 により索出され、この圧縮波形データが伸長部 6 により元の波形データへと復元され、検索部 5 を介して音響処理部 4 へと供給される。音響処理部 4 は、この波形データを音片編集部 8 へと供給する。

音片編集部 8 は、音響処理部 4 より波形データを返送されると、この波形データと、話速変換部 11 より供給された音片データのうち音片編集部 8 が選択したものとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する。

なお、話速変換部 11 より供給されたデータに欠落部分識別データが含まれていない場合は、第 1 の実施の形態と同様、音響処理部

4に波形の合成を指示することなく直ちに、音片編集部8が選択した音片データを、定型メッセージデータが示す定型メッセージ内の各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力すればよい。

以上説明したように、この第2の実施の形態の音声合成システムでは、音片データ同士の境界でのピッチ成分の周波数の不連続的な変化の量の累計が定型メッセージ全体で最小となるように音片データが選ばれ、録音編集方式により自然につながり合わせられるため、合成音声が自然なものとなる。また、この音声合成システムでは、処理が複雑な韻律予測は行われないので、簡単な構成で高速な処理にも追従できる。

なお、この第2の実施の形態の音声合成システムの構成も、上述のものに限られない。

例えば、ピッチ成分データは音片データが表す音片の先頭及び末尾でのピッチ長を表すデータであってもよい。この場合、音片編集部8は、話速変換部11より供給された各音片データの先頭及び末尾でのピッチ長を話速変換部11より供給されたピッチ成分データに基づいて特定し、定型メッセージ内で隣接する音片同士の境界でのピッチ長の差の絶対値を定型メッセージ全体で累計した値が最小になる、という条件を満たすように、音片データを選択すればよい。

また、音片編集部8は、例えば、言語処理部1と共にフリーテキストデータを取得し、このフリーテキストデータが表すフリーテキストに含まれる音片の波形とみなせる波形を表す音片データを、定型メッセージに含まれる音片の波形とみなせる波形を表す音片データを抽出する処理と実質的に同一の処理を行うことによって抽出し

て、音声の合成に用いてもよい。

この場合、音響処理部 4 は、音片編集部 8 が抽出した音片データが表す音片については、この音片の波形を表す波形データを検索部 5 に索出させなくてもよい。なお、音片編集部 8 は、音響処理部 4 が合成しなくてよい音片を音響処理部 4 に通知し、音響処理部 4 はこの通知に応答して、この音片を構成する単位音声の波形の検索を中止するようにすればよい。

また、音片編集部 8 は、例えば、音響処理部 4 と共に配信文字列データを取得し、この配信文字列データが表す配信文字列に含まれる音片の波形とみなせる波形を表す音片データを、定型メッセージに含まれる音片の波形とみなせる波形を表す音片データを抽出する処理と実質的に同一の処理を行うことによって抽出して、音声の合成に用いてもよい。この場合、音響処理部 4 は、音片編集部 8 が抽出した音片データが表す音片については、この音片の波形を表す波形データを検索部 5 に索出させなくてもよい。

### （第 3 の実施の形態）

次に、この発明の第 3 の実施の形態を説明する。この発明の第 3 の実施の形態に係る音声合成システムの物理的構成は、上述した第 1 の実施の形態における構成と実質的に同一である。

次に、この音声合成システムの動作を説明する。

この音声合成システムの言語処理部 1 がフリーテキストデータを外部から取得した場合、及び、音響処理部 4 が配信文字列データを取得した場合の動作は、第 1 又は第 2 の実施の形態の音声合成システムが行う動作と実質的に同一である。（なお、言語処理部 1 がフリーテキストデータを取得する手法や音響処理部 4 が配信文字列デ



ータを取得する手法はいずれも任意であり、例えば、いずれも第 1 又は第 2 の実施の形態における言語処理部 1 や音響処理部 4 が行う手法と同様の手法によりフリーテキストデータあるいは配信文字列データを取得すればよい。)

次に、音片編集部 8 が、定型メッセージデータ及び発声スピードデータを取得したとする。なお、音片編集部 8 が定型メッセージデータや発声スピードデータを取得する手法も任意であり、例えば、第 1 の実施の形態の音片編集部 8 が行う手法と同様の手法で定型メッセージデータや発声スピードデータを取得すればよい。あるいは、例えばこの音声合成システムがカーナビゲーションシステム等の車両内システムの一部をなすものであって、この車両内システムを構成するの他の装置（例えば、音声認識を行い、音声認識の結果得られた情報に基づいてエージェント処理を実行する装置など）が、ユーザーに対して発話する内容や発話スピードを決定し、決定結果を表すデータを生成するものである場合、この音声合成システムは、生成されたこのデータを受信（取得）し、定型メッセージデータ及び発声スピードデータとして扱うようにしてもよい。

定型メッセージデータ及び発声スピードデータが音片編集部 8 に供給されると、音片編集部 8 は、第 1 の実施の形態における音片編集部 8 と同様に、定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データをすべて索出するよう、検索部 9 に指示する。また、話速変換部 11 に対しても、第 1 の実施の形態における音片編集部 8 と同様に、話速変換部 11 に供給される音片データを変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致す

るようにすることを指示する。

すると、検索部 9、伸張部 6 及び話速変換部 11 が、第 1 の実施の形態における検索部 9、伸張部 6 及び話速変換部 11 の動作と実質的に同一の動作を行い、この結果、話速変換部 11 から音片編集部 8 へと、音片データ、音片読みデータ、この音片データが表す音片の発声スピードを表すスピード初期値データ及びピッチ成分データが供給される。また、欠落部分識別データが検索部 9 より話速変換部 11 へと供給された場合は、更にこの欠落部分識別データも音片編集部 8 へと供給される。

音片編集部 8 は、話速変換部 11 より音片データ、音片読みデータ及びピッチ成分データを供給されると、話速変換部 11 より供給された各々のピッチ成分データについて上述の値  $\alpha$ 、 $\beta$  の組及び／又は  $R_{max}$  を求め、また、このスピード初期値データと、音片編集部 8 に供給された定型メッセージデータ及び発声スピードデータとを用いて、上述の値  $d_t$  を求める。

そして、音片編集部 8 は、話速変換部 11 より供給されたそれぞれの音片データにつき、自ら求めた当該音片データ（以下、音片データ  $X$  と記す）についての  $\alpha$ 、 $\beta$ 、 $R_{max}$  及び  $d_t$  の値と、定型メッセージ内で当該音片データが表す音片の後に隣接する音片を表す音片データ（以下、音片データ  $Y$  と記す）のピッチ成分の周波数とに基づいて、数式 7 に示す評価値  $H_{XY}$  を特定する。

$$H_{XY} = (W_A \cdot \cos t\_A) + (W_B \cdot \cos t\_B) + (W_C \cdot \cos t\_C)$$

（ただし、 $W_A$ 、 $W_B$  及び  $W_C$  はいずれも所定の係数であり、 $W_A$  は 0 ではないものとする）

数式 7 の右辺に含まれる値  $\text{cost\_A}$  は、当該定型メッセージ内で互いに隣接する、音片データ X が表す音片と音片データ Y が表す音片との境界でのピッチ成分の周波数の差の絶対値の逆数である。

なお、音片編集部 8 は、 $\text{cost\_A}$  の値を特定するため、話速変換部 11 より供給されたピッチ成分データに基づき、話速変換部 11 より供給された各音片データの先頭及び末尾の各時点でのピッチ成分の周波数を特定するようにすればよい。

また、数式 7 の右辺に含まれる値  $\text{cost\_B}$  は、音片データ X について数式 8 に従って評価値  $\text{cost\_B}$  を求めた場合の値である。

$$\text{cost\_B} = 1 / (W_{B1} | 1 - \alpha | + W_{B2} | \beta | + W_{B3} \cdot dt)$$

(ただし、 $W_{B1}$ 、 $W_{B2}$  及び  $W_{B3}$  は所定の正の係数)

また、数式 7 の右辺に含まれる値  $\text{cost\_C}$  は、音片データ X について数式 9 に従って評価値  $\text{cost\_C}$  を求めた場合における値である。

$$\text{cost\_C} = 1 / (W_{C1} | R_{\max} | + W_{C2} \cdot dt)$$

(ただし、 $W_{C1}$  及び  $W_{C2}$  は所定の係数)

あるいは、音片編集部 8 は、数式 7 ～ 数式 9 に代えて、数式 10 及び数式 11 に従って評価値  $H_{XY}$  を特定するようにしてもよい。ただし、数式 10 に含まれる  $\text{cost\_B}$  及び  $\text{cost\_C}$  については、上述の係数  $W_{B3}$  及び  $W_{C3}$  の値はいずれも 0 とする。また、数式 8 及び数式 9 における  $(W_{B3} \cdot dt)$  及び  $(W_{C2} \cdot dt)$  の項を備えなくともよい。

$$H_{XY} = (W_A \cdot \text{cost\_A}) + (W_B \cdot \text{cost\_B}) + (W_C \cdot$$

$$\text{cost\_C}) + (W_D \cdot \text{cost\_D})$$

(ただし、 $W_D$ は0でない所定の係数)

$$\text{cost\_D} = 1 / (W_{d1} \cdot dt)$$

(ただし、 $W_{d1}$ は0でない所定の係数)

そして、音片編集部8は、話速変換部11より供給された各音片データのうちから、音片編集部8に供給された定型メッセージデータが表す定型メッセージを構成する音片1個につき1個ずつの音片データを選ぶことにより得られる各組み合わせのうち、組み合わせに属する各音片データの評価値 $H_{xy}$ の総和が最大となるものを、定型メッセージを読み上げる音声を合成するための最適な音片データの組み合わせとして選択する。

つまり、例えば第5図に示すように、定型メッセージデータが表す定型メッセージが音片A、B及びCより構成され、音片Aを表す音片データの候補として音片データA1、A2及びA3が索出され、音片Bを表す音片データの候補として音片データB1及びB2が索出され、音片Cを表す音片データの候補として音片データC1、C2及びC3が索出された場合、音片データA1、A2及びA3のうちから1個、音片データB1及びB2のうちから1個、音片データC1、C2及びC3のうちから1個、計3個選ぶことにより得られる組み合わせ計18通りのうち、組み合わせに属する各音片データの評価値 $H_{xy}$ の総和が最大となるものを、定型メッセージを読み上げる音声を合成するための最適な音片データの組み合わせとして選択する。

ただし、総和を求めるために用いられる評価値 $H_{xy}$ としては、組み合わせ内での音片の接続関係を正しく反映したものが選ばれるも

のとする。つまり、例えば組み合わせ内に、音片 p を表す音片データ P 及び音片 q を表す音片データ Q が含まれており、定型メッセージ内では音片 p が音片 q に先行する形で互いに隣接するという場合、音片データ P の評価値としては、音片 p が音片 q に先行する形で互いに隣接する場合における評価値  $H_{p,q}$  が用いられるものとする。

また、定型メッセージの末尾の音片（例えば、第 5 図を参照して前述した例でいえば、音片 C 1, C 2 及び C 3）については、後続する音片が存在しないため、 $cost\_A$  の値を定めることができない。このため、これら末尾の音片を表す音片データの評価値  $H_{x,y}$  を算定するにあたって、音片編集部 8 は、 $(W_A \cdot cost\_A)$  の値を 0 であるものとして扱い、一方、係数  $W_B$ ,  $W_C$  及び  $W_D$  の値は、それぞれ、他の音片データの評価値  $H_{x,y}$  を算定する場合とは異なる所定の値であるものとして扱う。

なお、音片編集部 8 は、数式 7 あるいは数式 11 を用いて、音片データ X について、当該音片データ X が表す音片の前に隣接する音片データ Y との関係を表す評価値を含むものとして評価値  $H_{x,y}$  を特定してもよい。この場合は、定型メッセージの先頭の音片について、先行する音片が存在しないため、 $cost\_A$  の値を定めることができないこととなる。このため、これら先頭の音片を表す音片データの評価値  $H_{x,y}$  を算定するにあたって、音片編集部 8 は、 $(W_A \cdot cost\_A)$  の値を 0 であるものとして扱い、一方、係数  $W_B$ ,  $W_C$  及び  $W_D$  の値は、それぞれ、他の音片データの評価値  $H_{x,y}$  を算定する場合とは異なる所定の値であるものとして扱うようにすればよい。

一方、音片編集部 8 は、話速変換部 11 より欠落部分識別データ

も供給されている場合には、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出して音響処理部 4 に供給し、この音片の波形を合成するよう指示する。

指示を受けた音響処理部 4 は、音片編集部 8 より供給された表音文字列を、配信文字列データが表す表音文字列と同様に扱う。この結果、この表音文字列に含まれる表音文字が示す音声の波形を表す圧縮波形データが検索部 5 により索出され、この圧縮波形データが伸長部 6 により元の波形データへと復元され、検索部 5 を介して音響処理部 4 へと供給される。音響処理部 4 は、この波形データを音片編集部 8 へと供給する。

音片編集部 8 は、音響処理部 4 より波形データを返送されると、この波形データと、話速変換部 11 より供給された音片データのうち、評価値  $H_{xy}$  の総和が最大となる組み合わせとして音片編集部 8 が選択した組み合わせに属するものとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する。

なお、話速変換部 11 より供給されたデータに欠落部分識別データが含まれていない場合は、第 1 の実施の形態と同様、音響処理部 4 に波形の合成を指示することなく直ちに、音片編集部 8 が選択した音片データを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力すればよい。

以上説明したように、この第 3 の実施の形態の音声合成システムでも、音片データが録音編集方式により自然につながり合わせられ、定型メッセージを読み上げる音声合成される。音片データペース

10の記憶容量は、音素毎に波形を記憶する場合に比べて小さくでき、また、高速に検索できる。このため、この音声合成システムは小型軽量に構成することができ、また高速な処理にも追従できる。

そして、第3の実施の形態の音声合成システムによれば、定型メッセージを読み上げる音声を合成するために選択される音片データの組み合わせの適切さを評価するための様々な評価基準（例えば、音片の波形の予測結果と音片データとの相関を1次回帰させた場合の勾配や切片による評価や、音片の時間差による評価や、音片データ同士の境界でのピッチ成分の周波数の不連続的な変化の量の累計、など）が、1個の評価値に影響を及ぼす形で総合的に反映され、この結果、最も自然な合成音声を合成するために選択すべき最適な音片データの組み合わせが、適正に決定される。

なお、この第3の実施の形態の音声合成システムの構成も、上述のものに限られない。

例えば、最適な音片データの組み合わせを選択するために音片編集部8が用いる評価値は数式7～13に示すものに限られず、音片データが表す音片を互いに結合して得られる音声、人の発する音声にどの程度類似又は相違しているかについての評価を表す任意の値であってよい。

また、評価値を表す数式（評価式）に含まれる変数ないし定数も必ずしも数式7～13に含まれているものに限られず、評価式としては、音片データが表す音片の特徴を示す任意のパラメータや、あるいは当該音片を互いに結合して得られる音声の特徴を示す任意のパラメータや、あるいは当該音声を人が発した場合に当該音声に備わると予測される特徴を示す任意のパラメータを含んだ数式が用い

られてよい。

また、最適な音片データの組み合わせを選択するための基準は必ずしも評価値の形で表現可能なものである必要はなく、音片データが表す音片を互いに結合して得られる音声の人が発する音声にどの程度類似又は相違しているかについての評価に基づいて音片データの最適な組み合わせを特定するに至るような基準である限り任意である。

また、音片編集部 8 は、例えば、言語処理部 1 と共にフリーテキストデータを取得し、このフリーテキストデータが表すフリーテキストに含まれる音片の波形とみなせる波形を表す音片データを、定型メッセージに含まれる音片の波形とみなせる波形を表す音片データを抽出する処理と実質的に同一の処理を行うことによって抽出して、音声の合成に用いてもよい。この場合、音響処理部 4 は、音片編集部 8 が抽出した音片データが表す音片については、この音片の波形を表す波形データを検索部 5 に索出させなくてもよい。なお、音片編集部 8 は、音響処理部 4 が合成しなくてよい音片を音響処理部 4 に通知し、音響処理部 4 はこの通知に応答して、この音片を構成する単位音声の波形の検索を中止するようにすればよい。

また、音片編集部 8 は、例えば、音響処理部 4 と共に配信文字列データを取得し、この配信文字列データが表す配信文字列に含まれる音片の波形とみなせる波形を表す音片データを、定型メッセージに含まれる音片の波形とみなせる波形を表す音片データを抽出する処理と実質的に同一の処理を行うことによって抽出して、音声の合成に用いてもよい。この場合、音響処理部 4 は、音片編集部 8 が抽出した音片データが表す音片については、この音片の波形を表す波



形データを検索部 5 に索出させなくてもよい。

以上、この発明の実施の形態を説明したが、この発明にかかる音声データ選択装置は、専用のシステムによらず、通常のコンピュータシステムを用いて実現可能である。

例えば、パーソナルコンピュータに上述の第 1 の実施の形態における言語処理部 1、一般単語辞書 2、ユーザ単語辞書 3、音響処理部 4、検索部 5、伸長部 6、波形データベース 7、音片編集部 8、検索部 9、音片データベース 10 及び話速変換部 11 の動作を実行させるためのプログラムを格納した媒体（CD-ROM、MO、フロッピー（登録商標）ディスク等）から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第 1 の実施の形態の本体ユニット M の機能を行わせることができる。

また、パーソナルコンピュータに、上述の第 1 の実施の形態における収録音片データセット記憶部 12、音片データベース作成部 13 及び圧縮部 14 の動作を実行させるためのプログラムを格納した媒体から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第 1 の実施の形態の音片登録ユニット R の機能を行わせることができる。

そして、これらのプログラムを実行し、第 1 の実施の形態における本体ユニット M や音片登録ユニット R として機能するパーソナルコンピュータが、第 1 図の音声合成システムの動作に相当する処理として、第 6 図～第 8 図に示す処理を行うものとする。

第 6 図は、このパーソナルコンピュータがフリーテキストデータを取得した場合の処理を示すフローチャートである。

第 7 図は、このパーソナルコンピュータが配信文字列データを取

得した場合の処理を示すフローチャートである。

第 8 図は、このパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

すなわち、まず、このパーソナルコンピュータが、外部より、上述のフリーテキストデータを取得すると（第 6 図、ステップ S 1 0 1）、このフリーテキストデータが表すフリーテキストに含まれるそれぞれの表意文字について、その読みを表す表音文字を、一般単語辞書 2 やユーザ単語辞書 3 を検索することにより特定し、この表意文字を、特定した表音文字へと置換する（ステップ S 1 0 2）。なお、このパーソナルコンピュータがフリーテキストデータを取得する手法は任意である。

そして、このパーソナルコンピュータは、フリーテキスト内の表意文字をすべて表音文字へと置換した結果を表す表音文字列が得られると、この表音文字列に含まれるそれぞれの表音文字について、当該表音文字が表す単位音声の波形を波形データベース 7 より検索し、表音文字列に含まれるそれぞれの表音文字が表す単位音声の波形を表す圧縮波形データを索出する（ステップ S 1 0 3）。

次に、このパーソナルコンピュータは、索出された圧縮波形データを、圧縮される前の波形データへと復元し（ステップ S 1 0 4）、復元された波形データを、表音文字列内での各表音文字の並びに従った順序で互いに結合し、合成音声データとして出力する（ステップ S 1 0 5）。なお、このパーソナルコンピュータが合成音声データを出力する手法は任意である。

また、このパーソナルコンピュータが、外部より、上述の配信文

字列データを任意の手法で取得すると(第7図、ステップS201)、この配信文字列データが表す表音文字列に含まれるそれぞれの表音文字について、当該表音文字が表す単位音声の波形を波形データベース7より検索し、表音文字列に含まれるそれぞれの表音文字が表す単位音声の波形を表す圧縮波形データを索出する(ステップS202)。

次に、このパーソナルコンピュータは、索出された圧縮波形データを、圧縮される前の波形データへと復元し(ステップS203)、復元された波形データを、表音文字列内での各表音文字の並びに従った順序で互いに結合し、合成音声データとしてステップS105の処理と同様の処理により出力する(ステップS204)。

一方、このパーソナルコンピュータが、外部より、上述の定型メッセージデータ及び発声スピードデータを任意の手法により取得すると(第8図、ステップS301)、まず、この定型メッセージデータが表す定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データをすべて索出する(ステップS302)。

また、ステップS302では、該当する圧縮音片データに対応付けられている上述の音片読みデータ、スピード初期値データ及びピッチ成分データも索出する。なお、1個の音片につき複数の圧縮音片データが該当する場合は、該当する圧縮音片データすべてを索出する。一方、圧縮音片データを索出できなかった音片があった場合は、上述の欠落部分識別データを生成する。

次に、このパーソナルコンピュータは、索出された圧縮音片データを、圧縮される前の音片データへと復元する(ステップS303)。

そして、復元された音片データを、上述の音片編集部 8 が行う処理と同様の処理により変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致させる（ステップ S 3 0 4）。なお、発声スピードデータが供給されていない場合は、復元された音片データを変換しなくてもよい。

次に、このパーソナルコンピュータは、音片の時間長が変換された音片データのうちから、定型メッセージを構成する音片の波形に最も近い波形を表す音片データを、上述の音片編集部 8 が行う処理と同様の処理を行うことにより、音片 1 個につき 1 個ずつ選択する（ステップ S 3 0 5 ～ S 3 0 8）。

すなわち、このパーソナルコンピュータは、定型メッセージデータが表す定型メッセージに韻律予測の手法に基づいた解析を加えることにより、この定型メッセージの韻律を予測する（ステップ S 3 0 5）。そして、定型メッセージ内のそれぞれの音片について、この音片のピッチ成分の周波数の時間変化の予測結果と、この音片と読みが合致する音片の波形を表す音片データのピッチ成分の周波数の時間変化を表すピッチ成分データとの相関を求める（ステップ S 3 0 6）。より具体的には、索出された各々のピッチ成分データについて、例えば、上述した勾配  $\alpha$  及び切片  $\beta$  の値を求める。

一方で、このパーソナルコンピュータは、索出されたスピード初期値データと、外部より取得した定型メッセージデータ及び発声スピードデータとを用いて、上述の値  $d$   $t$  を求める（ステップ S 3 0 7）。

そして、このパーソナルコンピュータは、ステップ S 3 0 6 で求めた  $\alpha$ 、 $\beta$  の値、及び、ステップ S 3 0 7 で求めた  $d$   $t$  の値に基づ

いて、定型メッセージ内の音片の読みと一致する音片を表す音片データのうち、上述の評価値  $cost1$  が最大となるものを選択する（ステップ S 3 0 8）。

なお、このパーソナルコンピュータは、ステップ S 3 0 6 で、上述の  $\alpha$  及び  $\beta$  の値を求める代わりに、上述の  $R_{xy}(j)$  の最大値を求めるようにしてもよい。この場合は、ステップ S 3 0 8 で、 $R_{xy}(j)$  の最大値と、ステップ S 3 0 7 で求めた係数  $dt$  とに基づいて、定型メッセージ内の音片の読みと一致する音片を表す音片データのうち、上述の評価値  $cost2$  が最大となるものを選択すればよい。

一方、このパーソナルコンピュータは、欠落部分識別データを生成した場合、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出し、この表音文字列につき、音素毎に、配信文字列データが表す表音文字列と同様に扱って上述のステップ S 2 0 2 ～ S 2 0 3 の処理を行うことにより、この表音文字列内の各表音文字が示す音声の波形を表す波形データを復元する（ステップ S 3 0 9）。

そして、このパーソナルコンピュータは、復元した波形データと、ステップ S 3 0 8 で選択した音片データとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する（ステップ S 3 1 0）。

また、パーソナルコンピュータに上述の第 2 の実施の形態における言語処理部 1、一般単語辞書 2、ユーザ単語辞書 3、音響処理部 4、検索部 5、伸長部 6、波形データベース 7、音片編集部 8、検索部 9、音片データベース 10 及び話速変換部 11 の動作を実行さ

せるためのプログラムを格納した媒体から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第2の実施の形態における本体ユニットMの機能を行わせることができる。

また、パーソナルコンピュータに上述の第2の実施の形態における収録音片データセット記憶部12、音片データベース作成部13及び圧縮部14の動作を実行させるためのプログラムを格納した媒体から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第2の実施の形態における音片登録ユニットRの機能を行わせることができる。

そして、これらのプログラムを実行し、第2の実施の形態における本体ユニットMや音片登録ユニットRとして機能するパーソナルコンピュータが、第1図の音声合成システムの動作に相当する処理として、第6図及び第7図に示す上述の処理を行い、また、第9図に示す処理を行うものとする。

第9図は、このパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

すなわち、このパーソナルコンピュータが、外部より、上述の定型メッセージデータ及び発声スピードデータを任意の手法により取得すると（第9図、ステップS401）、まず、上述のステップS302の処理と同様に、この定型メッセージデータが表す定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データと、該当する圧縮音片データに対応付けられている上述の音片読みデータ、スピード初期値データ

及びピッチ成分データとを、すべて索出する(ステップS 4 0 2)。  
なお、ステップS 4 0 2でも、1個の音片につき複数の圧縮音片データが該当する場合は該当する圧縮音片データすべてを索出し、一方で圧縮音片データを索出できなかった音片があった場合は、上述の欠落部分識別データを生成する。

次に、このパーソナルコンピュータは、索出された圧縮音片データを、圧縮される前の音片データへと復元し(ステップS 4 0 3)、復元された音片データを、上述の音片編集部8が行う処理と同様の処理により変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致させる(ステップS 4 0 4)。  
なお、発声スピードデータが供給されていない場合は、復元された音片データを変換しなくてもよい。

次に、このパーソナルコンピュータは、音片の時間長が変換された音片データのうちから、定型メッセージを構成する音片の波形とみなせる波形を表す音片データを、上述の第2の実施の形態における音片編集部8が行う処理と同様の処理を行うことにより、音片1個につき1個ずつ選択する(ステップS 4 0 5～S 4 0 6)。

具体的には、まず、このパーソナルコンピュータは、音片の時間長が変換された各音片データの先頭及び末尾の各時点でのピッチ成分の周波数を、索出されたピッチ成分データに基づいて特定する(ステップS 4 0 5)。そして、これらの音片データのうちから、定型メッセージ内で隣接する音片同士の境界でのピッチ成分の周波数の差の絶対値を定型メッセージ全体で累計した値が最小になる、という条件を満たすように、音片データを選択する(ステップS 4 0 6)。この条件を満たす音片データを選択するために、このパーソナルコ

ンピュータは、例えば、定型メッセージ内で隣接する音片同士の境界でのピッチ成分の周波数の差の絶対値を距離として定義し、DPマッチングの手法により音片データを選ぶようにすればよい。

一方、このパーソナルコンピュータは、欠落部分識別データを生成した場合、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出し、この表音文字列につき、音素毎に、配信文字列データが表す表音文字列と同様に扱って上述のステップS202～S203の処理を行うことにより、この表音文字列内の各表音文字が示す音声の波形を表す波形データを復元する（ステップS407）。

そして、このパーソナルコンピュータは、復元した波形データと、ステップS406で選択した音片データとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する（ステップS408）。

また、パーソナルコンピュータに上述の第3の実施の形態における言語処理部1、一般単語辞書2、ユーザ単語辞書3、音響処理部4、検索部5、伸長部6、波形データベース7、音片編集部8、検索部9、音片データベース10及び話速変換部11の動作を実行させるためのプログラムを格納した媒体から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第3の実施の形態における本体ユニットMの機能を行わせることができる。

また、パーソナルコンピュータに上述の第3の実施の形態における収録音片データセット記憶部12、音片データベース作成部13及び圧縮部14の動作を実行させるためのプログラムを格納した媒



体から該プログラムをインストールすることにより、当該パーソナルコンピュータに、上述の第3の実施の形態における音片登録ユニットRの機能を行わせることができる。

そして、これらのプログラムを実行し、第3の実施の形態における本体ユニットMや音片登録ユニットRとして機能するパーソナルコンピュータが、第1図の音声合成システムの動作に相当する処理として、第6図及び第7図に示す上述の処理を行い、また、第10図に示す処理を行うものとする。

第10図は、このパーソナルコンピュータが定型メッセージデータ及び発声スピードデータを取得した場合の処理を示すフローチャートである。

すなわち、このパーソナルコンピュータが、外部より、上述の定型メッセージデータ及び発声スピードデータを任意の手法により取得すると（第10図、ステップS501）、まず、上述のステップS302の処理と同様に、この定型メッセージデータが表す定型メッセージに含まれる音片の読みを表す表音文字に合致する表音文字が対応付けられている圧縮音片データと、該当する圧縮音片データに対応付けられている上述の音片読みデータ、スピード初期値データ及びピッチ成分データとを、すべて索出する（ステップS502）。なお、ステップS502でも、1個の音片につき複数の圧縮音片データが該当する場合は該当する圧縮音片データすべてを索出し、一方で圧縮音片データを索出できなかった音片があった場合は、上述の欠落部分識別データを生成する。

次に、このパーソナルコンピュータは、索出された圧縮音片データを、圧縮される前の音片データへと復元し（ステップS503）、

復元された音片データを、上述の音片編集部 8 が行う処理と同様の処理により変換して、当該音片データが表す音片の時間長を、発声スピードデータが示すスピードに合致させる(ステップ S 5 0 4)。なお、発声スピードデータが供給されていない場合は、復元された音片データを変換しなくてもよい。

次に、このパーソナルコンピュータは、音片の時間長が変換された音片データのうちから、定型メッセージを読み上げる音声を合成するための最適な音片データの組み合わせを、上述の第 3 の実施の形態における音片編集部 8 が行う処理と同様の処理を行うことにより選択する(ステップ S 5 0 5 ~ S 5 0 7)。

すなわち、まず、このパーソナルコンピュータは、ステップ S 5 0 2 で索出された各々のピッチ成分データについて上述の値  $\alpha$ 、 $\beta$  の組及び／又は  $R_{max}$  を求め、また、このスピード初期値データと、ステップ S 5 0 1 で取得した定型メッセージデータ及び発声スピードデータとを用いて、上述の値  $d_t$  を求める(ステップ S 5 0 5)。

次に、このパーソナルコンピュータは、ステップ S 5 0 4 で変換されたそれぞれの音片データにつき、ステップ S 5 0 5 で求めた  $\alpha$ 、 $\beta$ 、 $R_{max}$  及び  $d_t$  の値と、定型メッセージ内で当該音片データが表す音片の後に隣接する音片を表す音片データのピッチ成分の周波数とに基づいて、上述した評価値  $H_{xy}$  を特定する(ステップ S 5 0 6)。

そして、このパーソナルコンピュータは、ステップ S 5 0 4 で変換された各音片データのうちから、ステップ S 5 0 1 で取得した定型メッセージデータが表す定型メッセージを構成する音片 1 個につき

1個ずつの音片データを選ぶことにより得られる各組み合わせのうち、組み合わせに属する各音片データの評価値 $H_{x,y}$ の総和が最大となるものを、定型メッセージを読み上げる音声を合成するための最適な音片データの組み合わせとして選択する(ステップS507)。ただし、総和を求めるために用いられる評価値 $H_{x,y}$ としては、組み合わせ内での音片の接続関係を正しく反映したものが選ばれるものとする。

一方、このパーソナルコンピュータは、欠落部分識別データを生成した場合、欠落部分識別データが示す音片の読みを表す表音文字列を定型メッセージデータより抽出し、この表音文字列につき、音素毎に、配信文字列データが表す表音文字列と同様に扱って上述のステップS202～S203の処理を行うことにより、この表音文字列内の各表音文字が示す音声の波形を表す波形データを復元する(ステップS508)。

そして、このパーソナルコンピュータは、復元した波形データと、ステップS507で選択した組み合わせに属する音片データとを、定型メッセージデータが示す定型メッセージ内での各音片の並びに従った順序で互いに結合し、合成音声を表すデータとして出力する(ステップS509)。

なお、パーソナルコンピュータに本体ユニットMや音片登録ユニットRの機能を行わせるプログラムは、例えば、通信回線の掲示板(BBS)にアップロードし、これを通信回線を介して配信してもよく、また、これらのプログラムを表す信号により搬送波を変調し、得られた変調波を伝送し、この変調波を受信した装置が変調波を復調してこれらのプログラムを復元するようにしてもよい。

そして、これらのプログラムを起動し、OSの制御下に、他のアプリケーションプログラムと同様に実行することにより、上述の処理を実行することができる。

なお、OSが処理の一部を分担する場合、あるいは、OSが本願発明の1つの構成要素の一部を構成するような場合には、記録媒体には、その部分を除いたプログラムを格納してもよい。この場合も、この発明では、その記録媒体には、コンピュータが実行する各機能又はステップを実行するためのプログラムが格納されているものとする。

#### 産業上利用可能性

本発明によれば、簡単な構成で高速に自然な合成音声を得るための音声選択装置、音声選択方法及びプログラムが実現される。

## 請求の範囲

1. 音声の波形を表す音声データを複数記憶する記憶手段と、  
文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出する検索手段と、

索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する選択手段と、

より構成されることを特徴とする音声データ選択装置。

2. 選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を更に備える、

ことを特徴とする請求項1に記載の音声データ選択装置。

3. 音声の波形を表す音声データを複数記憶し、  
文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出し、

索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する、

ことを特徴とする音声データ選択方法。

4. コンピュータを、

音声の波形を表す音声データを複数記憶する記憶手段と、

文章を表す文章情報を入力し、各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表している音声データを索出する検索手段と、

索出された音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ、互いに隣接する音片同士の境界でのピッチの差を前記文章全体で累計した値が最小となるように選択する選択手段と、

して機能させるためのプログラム。

5. 音声の波形を表す音声データを複数記憶する記憶手段と、

文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測する予測手段と、

各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する選択手段と、

より構成されることを特徴とする音声選択装置。

6. 前記選択手段は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での1次回帰を行う回帰計算の結果に基づいて、当該音声データのピッチの時間変化と前記予測手段による予測の結果との相関の強さを特定するものである、

ことを特徴とする請求項5に記載の音声選択装置。

7. 前記選択手段は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変

化との間の相関係数に基づいて、当該音声データのピッチの時間変化と前記予測手段による予測の結果との相関の強さを特定するものである、

ことを特徴とする請求項 5 に記載の音声選択装置。

8. 音声の波形を表す音声データを複数記憶する記憶手段と、  
文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測する予測手段と、

前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択する選択手段と、より構成されており、

前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数より得られるものである、

ことを特徴とする音声選択装置。

9. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での 1 次回帰により得られる 1 次関数の勾配からなる、

ことを特徴とする請求項 8 に記載の音声選択装置。

10. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での 1 次回帰により得られる 1 次関数の切片からなる、

る、

ことを特徴とする請求項 8 に記載の音声選択装置。

1 1. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との間の相関係数からなる、

ことを特徴とする請求項 8 に記載の音声選択装置。

1 2. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化を表すデータを種々のビット数循環シフトしたものが表す関数と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果を表す関数との相関係数の最大値からなる、

ことを特徴とする請求項 8 に記載の音声選択装置。

1 3. 前記記憶手段は、音声データの読みを表す表音データを、当該音声データに対応付けて記憶しており、

前記選択手段は、前記文章内の音片の読みに合致する読みを表す表音データが対応付けられている音声データを、当該音片と読みが共通する音片の波形を表す音声データとして扱う、

ことを特徴とする請求項 5 乃至 1 2 のいずれか 1 項に記載の音声選択装置。

1 4. 選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を更に備える、

ことを特徴とする請求項 5 乃至 1 3 のいずれか 1 項に記載の音声選択装置。

1 5. 前記文章内の音片のうち、前記選択手段が音声データを選択できなかった音片について、前記記憶手段が記憶する音声データを用いることなく、当該音片の波形を表す音声データを合成する欠落



部分合成手段を備え、

前記音声合成手段は、前記選択手段が選択した音声データ及び前記欠落部分合成手段が合成した音声データを互いに結合することにより、合成音声を表すデータを生成する、

ことを特徴とする請求項 14 に記載の音声選択装置。

16. 音声の波形を表す音声データを複数記憶し、

文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測し、

各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する、

ことを特徴とする音声選択方法。

17. 音声の波形を表す音声データを複数記憶し、

文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測し、

前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択する、ことを特徴とする音声選択方法であって、

前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間

長の予測結果との差の関数より得られるものである、

ことを特徴とする音声選択方法。

18. コンピュータを、

音声の波形を表す音声データを複数記憶する記憶手段と、

文章を表す文章情報を入力し、当該文章を構成する音片について韻律予測を行うことにより、当該音片のピッチの時間変化を予測する予測手段と、

各前記音声データのうちから、前記文章を構成する音片と読みが共通する音片の波形を表していて、且つ、ピッチの時間変化が前記予測手段による予測の結果と最も高い相関を示す音声データを選択する選択手段と、

して機能させるためのプログラム。

19. コンピュータを、

音声の波形を表す音声データを複数記憶する記憶手段と、

文章を表す文章情報を入力し、当該文章内の音片について韻律予測を行うことにより、当該音片の時間長、及び、当該音片のピッチの時間変化を予測する予測手段と、

前記文章内の音片と読みが共通する音片の波形を表す各々の音声データについての評価値を特定し、評価値が最も高い評価を表している音声データを選択する選択手段として機能させるためのプログラムであって、

前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間

長の予測結果との差の関数より得られるものである、

ことを特徴とするプログラム。

20. 音声の波形を表す音声データを複数記憶する記憶手段と、  
文章を表す文章情報を入力する文章情報入力手段と、

前記文章情報が表す文章内の音片と読みが共通する部分を有する  
音声データを索出する検索部と、前記索出されたそれぞれの音声デ  
ータを文章情報が表す文章に従って接続した際に互いに隣接する音  
声データ同士の関係に基づいた所定の評価基準に従って評価値を求  
め、出力する音声データの組み合わせを当該評価値に基づいて選択  
する選択手段と、を備える、

ことを特徴とする音声データ選択装置。

21. 前記評価基準は、互いに隣接する音声データ同士の関係を示  
す評価値を定める基準であって、前記評価値は、前記音声データが  
表す音声の特徴を示すパラメータ、前記音声データが表す音声を互  
いに結合して得られる音声の特徴を示すパラメータ、及び、発話時  
間長に関する特徴を示すパラメータのうち、少なくともいずれかを  
含む評価式に基づいて得られるものである、

ことを特徴とする請求項20に記載の音声データ選択装置。

22. 前記評価基準は、互いに隣接する音声データ同士の関係を示  
す評価値を定める基準であって、前記評価値は、前記音声データが  
表す音声を互いに結合して得られる音声の特徴を示すパラメータを  
含み、また、前記音声データが表す音声の特徴を示すパラメータと  
発話時間長に関する特徴を示すパラメータのうち、少なくともいずれ  
れかを含む評価式に基づいて得られるものである、

ことを特徴とする請求項20に記載の音声データ選択装置。

23. 前記音声データが表す音声とを互いに結合して得られる音声の特徴を示すパラメータは、前記文章情報が表す文章内の音片と読みが共通する部分を有する音声の波形を表す音声データのうちから、前記文章を構成するそれぞれの音片に相当する音声データを1個ずつ選択した場合における、互いに隣接する音声データ同士の境界でのピッチの差に基づいて得られるものである、

ことを特徴とする請求項21又は22に記載の音声データ選択装置。

24. 前記評価基準は、更に音声データが表す音声との韻律予測結果との相関ないし差異を示す評価値を定める基準を含み、前記評価値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との相関を表す数値の関数、及び／又は、当該音声データが表す音片の時間長と、当該音片と読みが共通する前記文章内の音片の時間長の予測結果との差の関数に基づいて得られるものである、

ことを特徴とする請求項20乃至23のいずれか1項に記載の音声データ選択装置。

25. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化との間での1次回帰により得られる1次関数の勾配及び／又は切片からなる、

ことを特徴とする請求項24に記載の音声データ選択装置。

26. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果との間の相関係数からなる、

ことを特徴とする請求項 24 又は 25 に記載の音声データ選択装置。

27. 前記相関を表す数値は、音声データが表す音片のピッチの時間変化を表すデータを種々のビット数循環シフトしたものが表す関数と、当該音片と読みが共通する前記文章内の音片のピッチの時間変化の予測結果を表す関数との相関係数の最大値からなる、

ことを特徴とする請求項 24 又は 25 に記載の音声データ選択装置。

28. 前記記憶手段は、音声データの読みを表す表音データを、当該音声データに対応付けて記憶しており、

前記選択手段は、前記文章内の音片の読みに合致する読みを表す表音データが対応付けられている音声データを、当該音片と読みが共通する音片の波形を表す音声データとして扱う、

ことを特徴とする請求項 20 乃至 27 のいずれか 1 項に記載の音声データ選択装置。

29. 選択された音声データを互いに結合することにより、合成音声を表すデータを生成する音声合成手段を更に備える、

ことを特徴とする請求項 20 乃至 28 のいずれか 1 項に記載の音声データ選択装置。

30. 前記文章内の音片のうち、前記選択手段が音声データを選択できなかった音片について、前記記憶手段が記憶する音声データを用いることなく、当該音片の波形を表す音声データを合成する欠落部分合成手段を備え、

前記音声合成手段は、前記選択手段が選択した音声データ及び前記欠落部分合成手段が合成した音声データを互いに結合することに

より、合成音声を表すデータを生成する、

ことを特徴とする請求項 29 に記載の音声データ選択装置。

31. 音声の波形を表す音声データを複数記憶し、

文章を表す文章情報を入力し、

前記文章情報が表す文章内の音片と読みが共通する部分を有する音声データを索出し、

前記索出されたそれぞれの音声データを文章情報が表す文章に従って接続した際に互いに隣接する音声データ同士の関係に基づいた所定の評価基準に従って評価値を求め、出力する音声データの組み合わせを当該評価値に基づいて選択する、

ことを特徴とする音声データ選択方法。

32. コンピュータを、

音声の波形を表す音声データを複数記憶する記憶手段と、

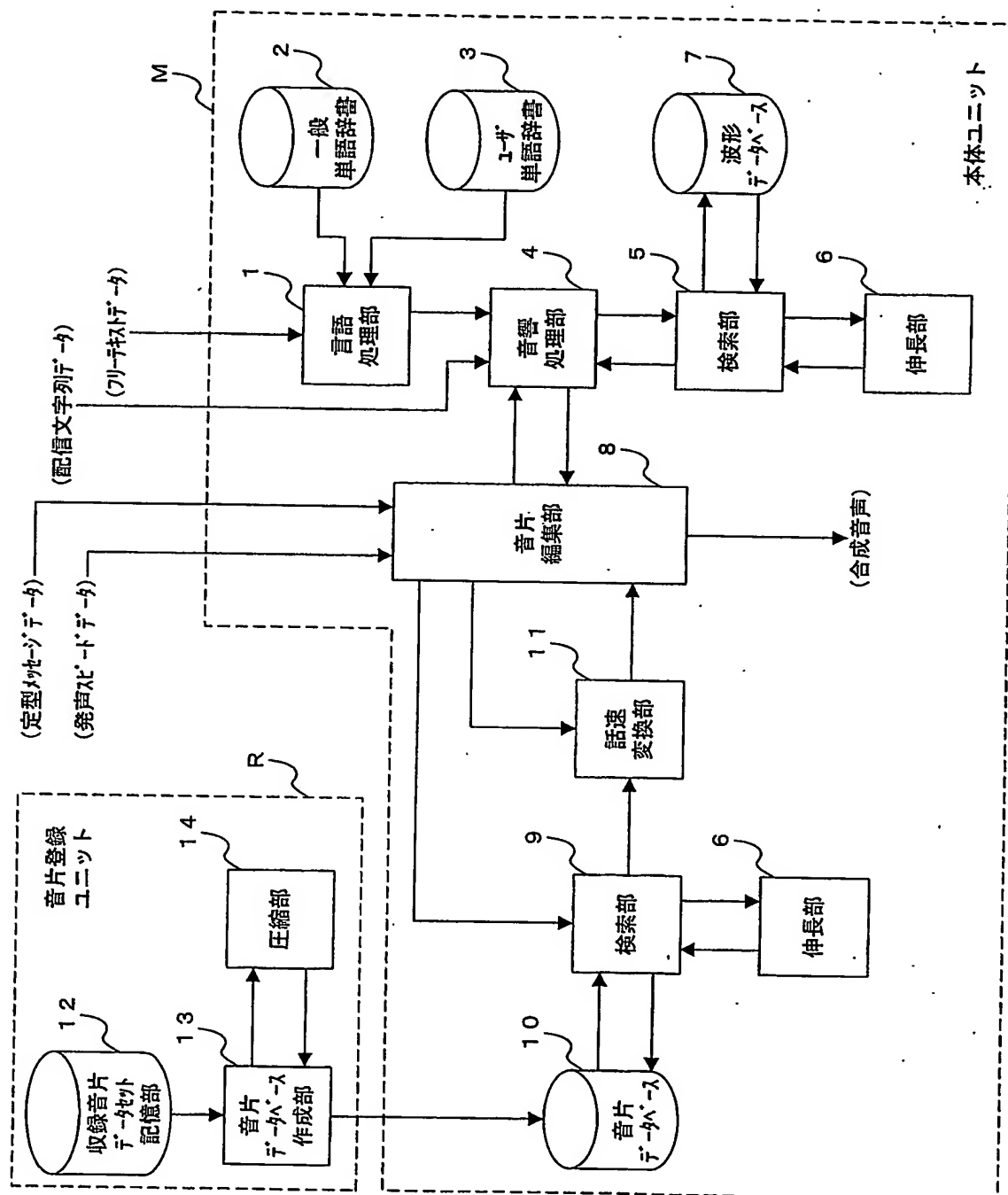
文章を表す文章情報を入力する文章情報入力手段と、

前記文章情報が表す文章内の音片と読みが共通する部分を有する音声データを索出する検索部と、

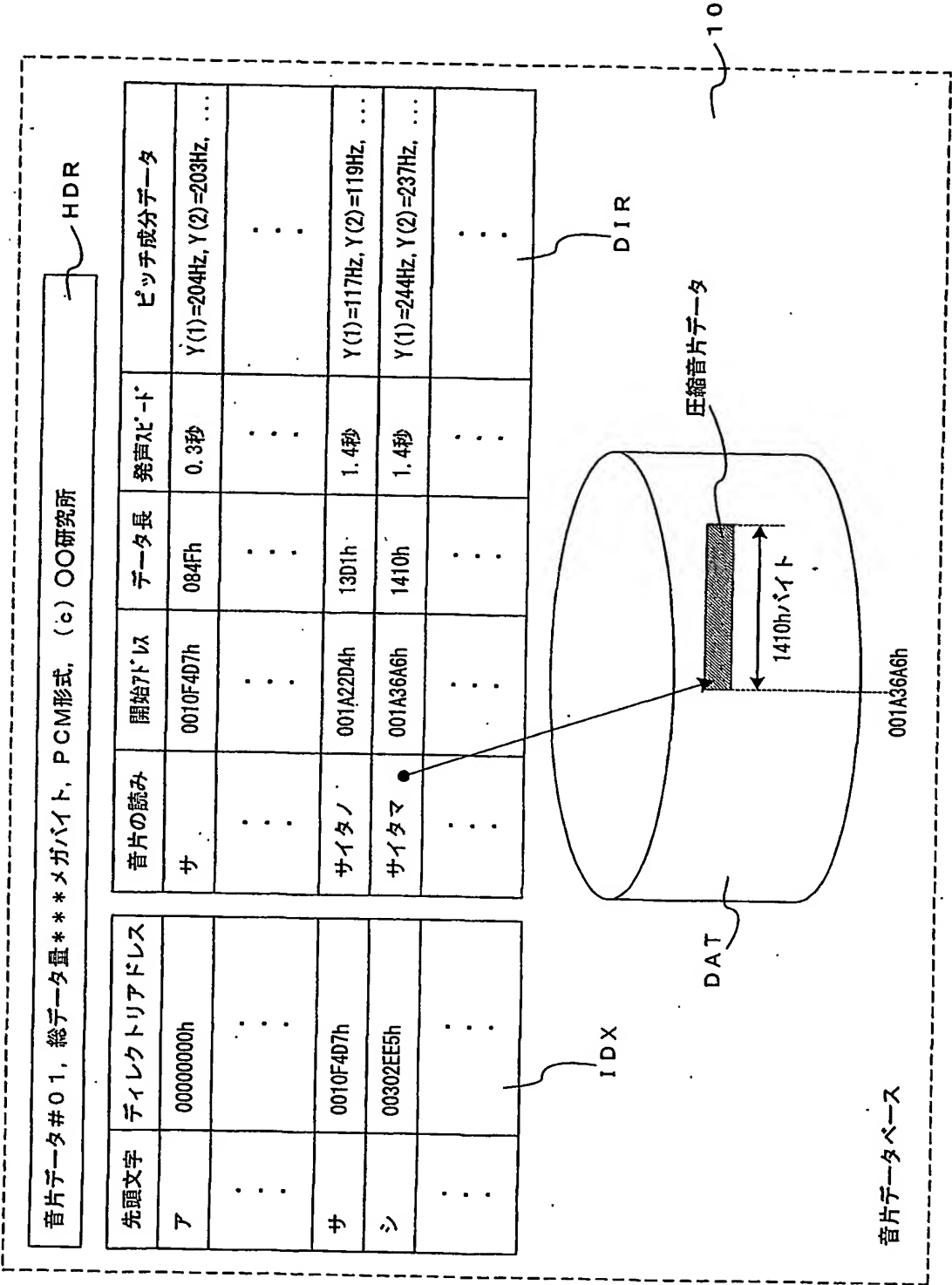
前記索出されたそれぞれの音声データを文章情報が表す文章に従って接続した際に互いに隣接する音声データ同士の関係に基づいた所定の評価基準に従って評価値を求め、出力する音声データの組み合わせを当該評価値に基づいて選択する選択手段と、

して機能させるためのプログラム。

第1図



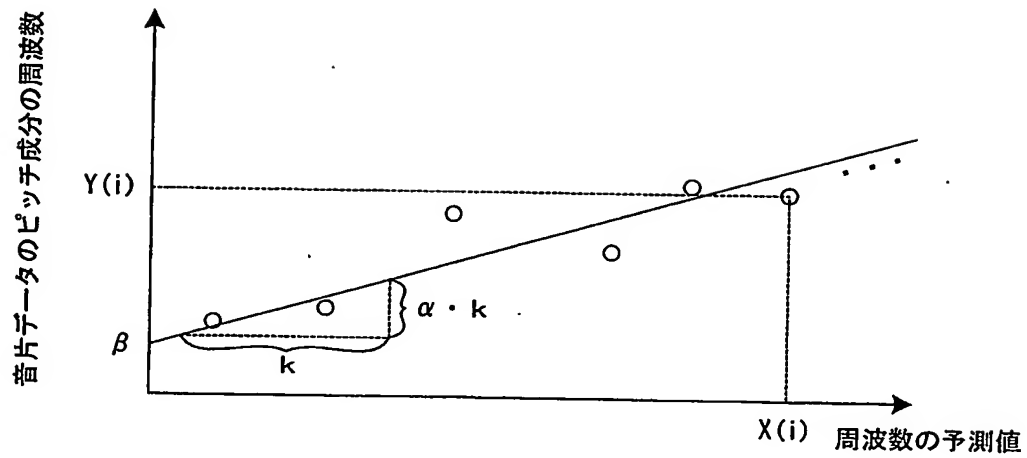
第2図



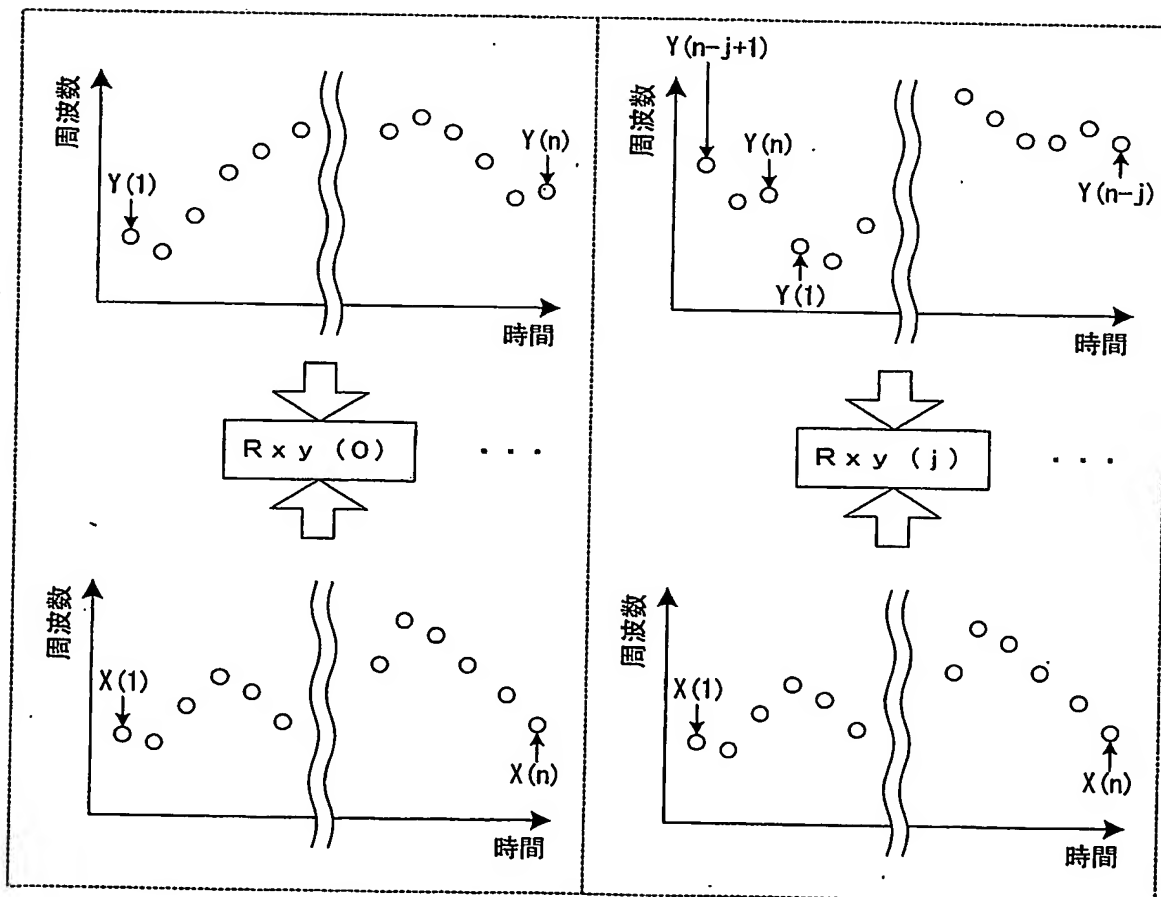


第3図

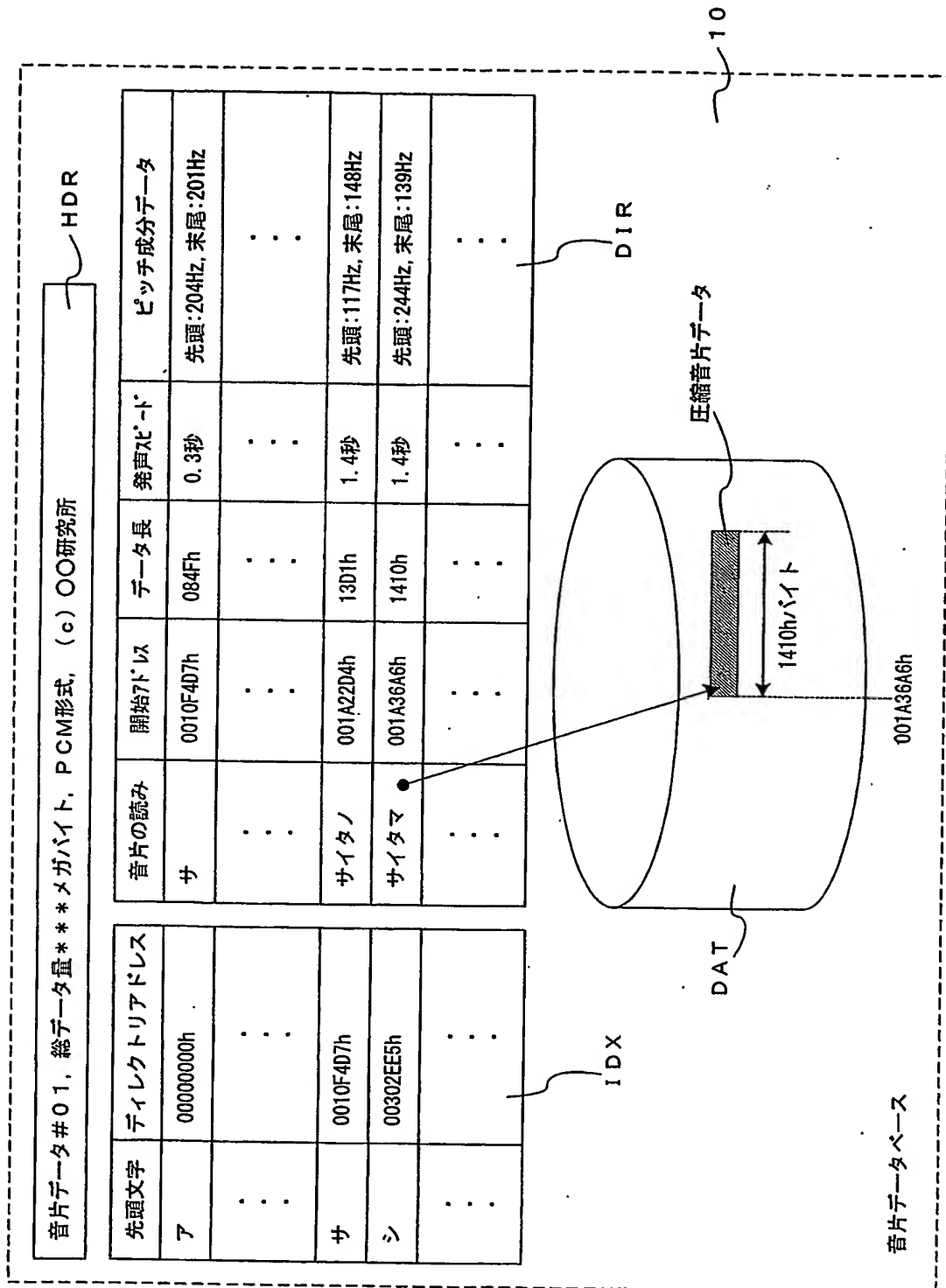
(a)



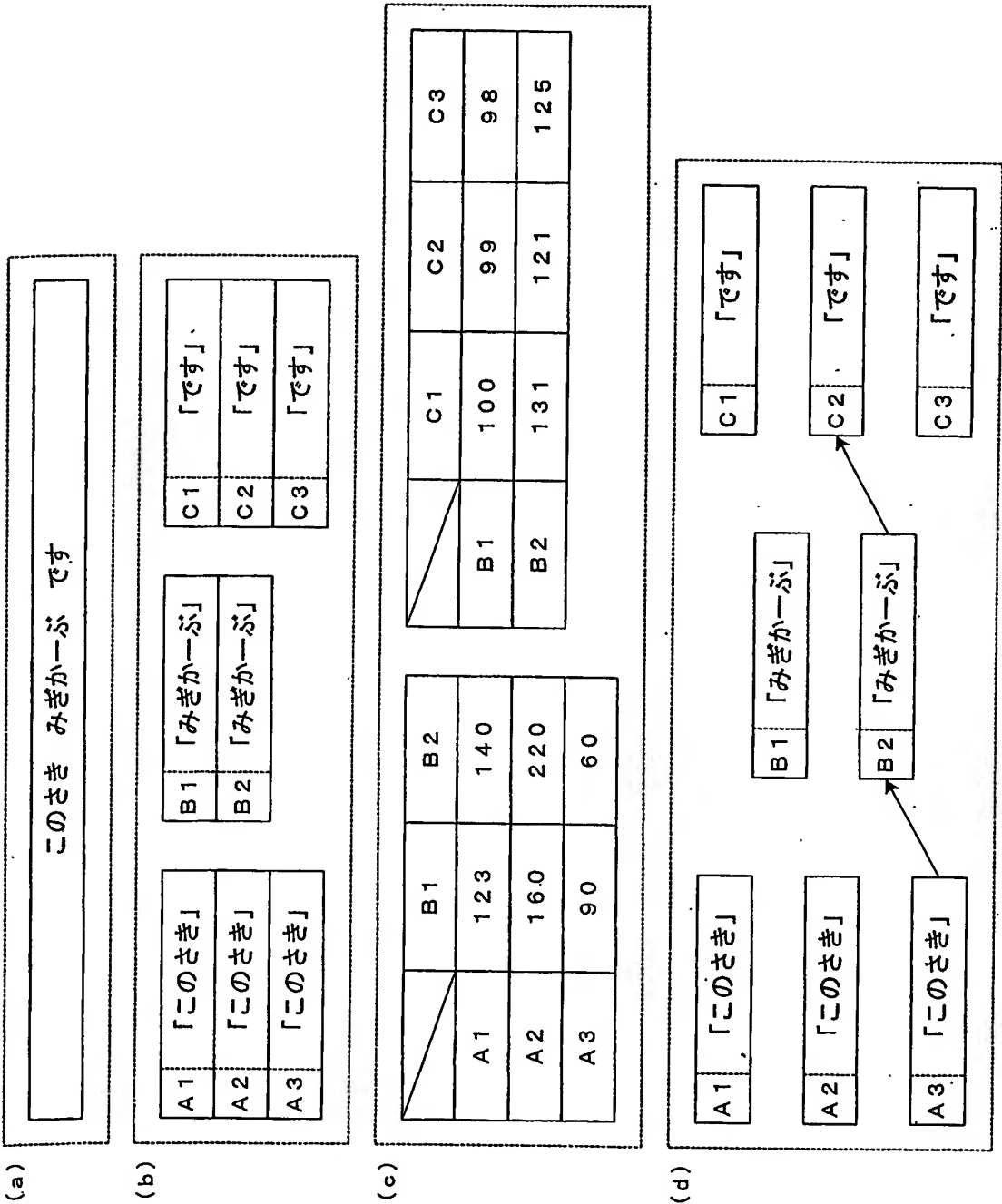
(b)



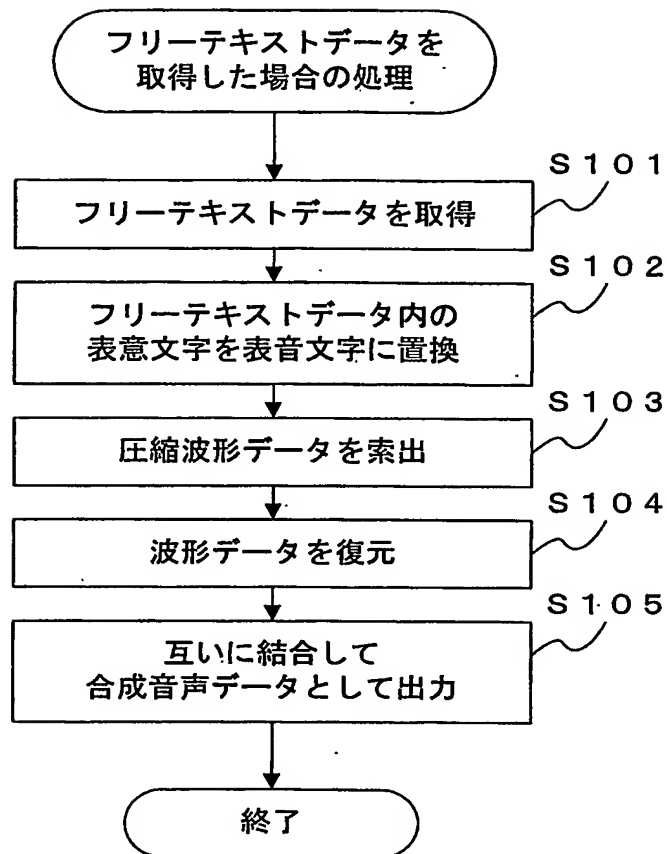
第4図



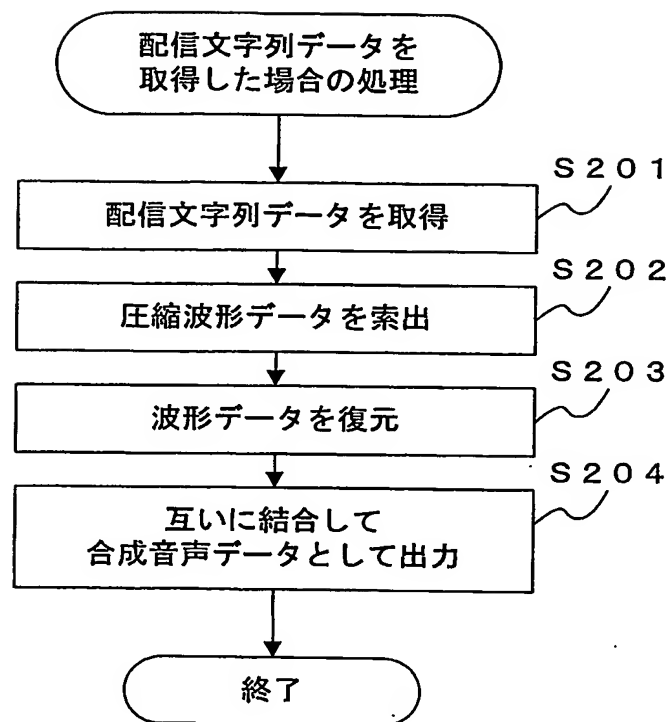
第5図



第6図

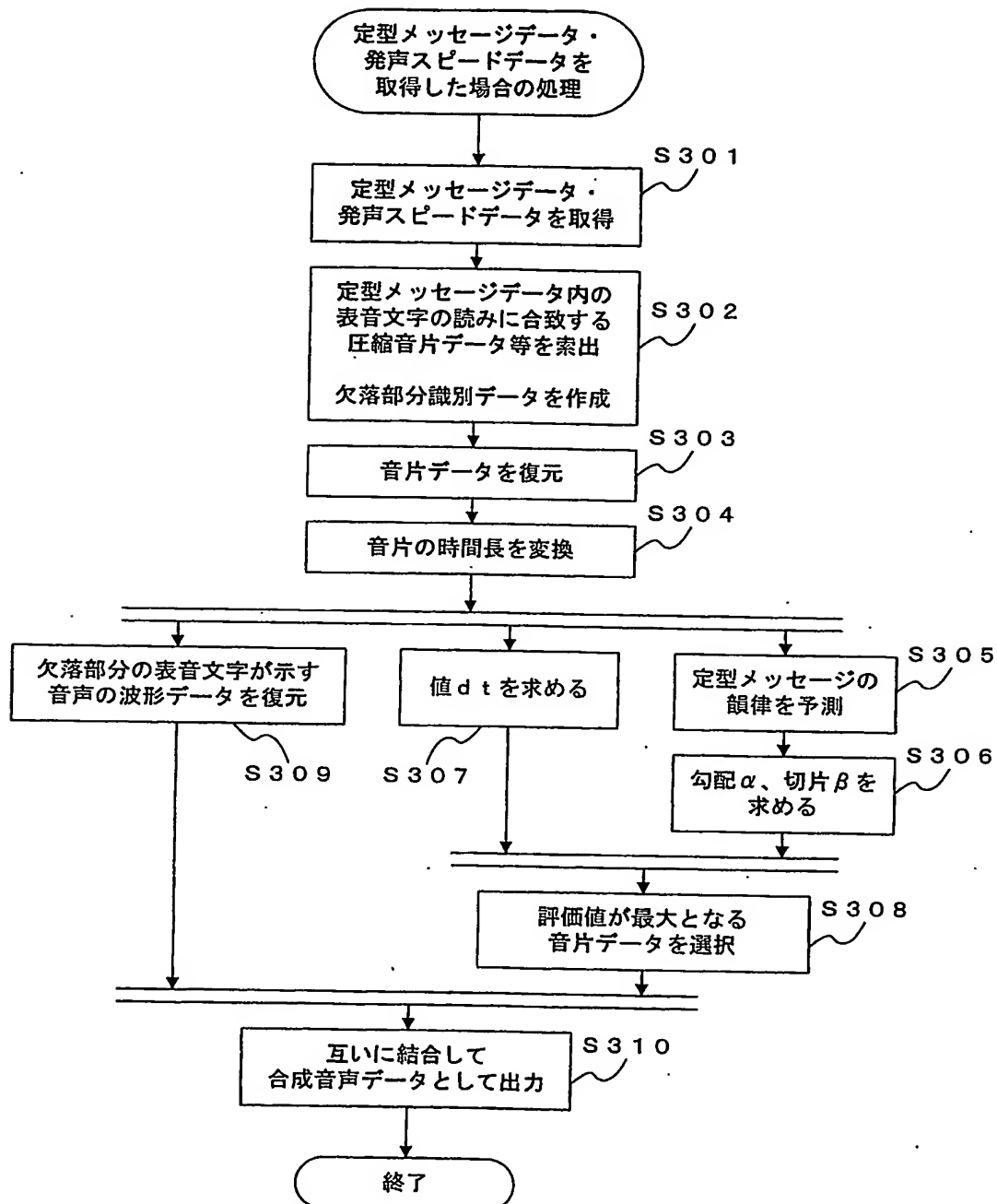


第7図



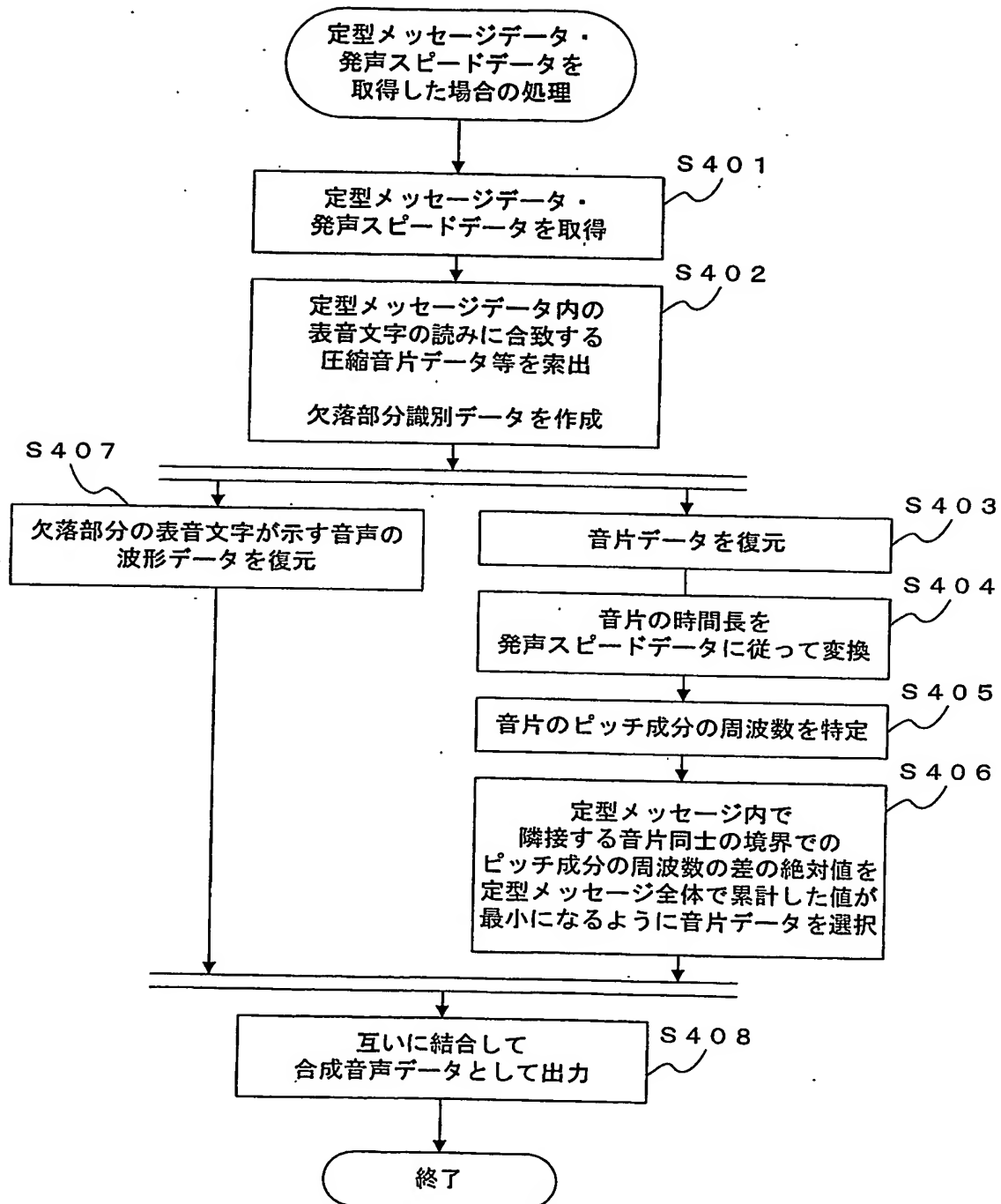
8/10

第8図



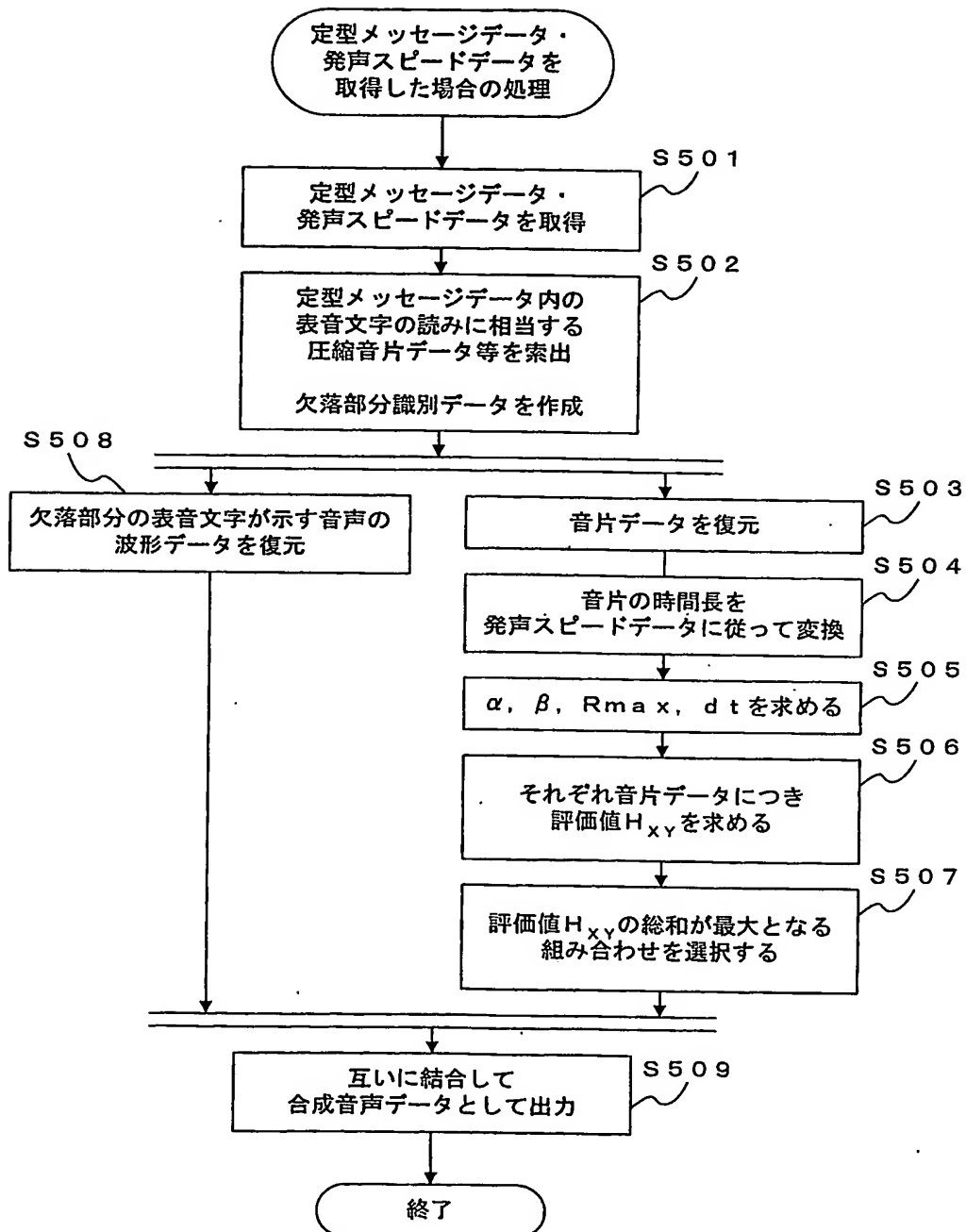
9/10

第9図



10/10

第10図





# INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/008088

## A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl<sup>7</sup> G10L13/06

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
Int.Cl<sup>7</sup> G10L13/00-13/08

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2004
Kokai Jitsuyo Shinan Koho	1971-2004	Toroku Jitsuyo Shinan Koho	1994-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2003-513311 A (Siemens AG.), 08 April, 2003 (08.04.03), Full text; Figs. 1 to 6 & WO 01/31434 A2 & EP 1224531 A2	1-32
A	JP 9-44191 A (Sanyo Electric Co., Ltd.), 14 February, 1997 (14.02.97), Full text; Figs. 1 to 9 (Family: none)	1-32
A	JP 10-97268 A (Sanyo Electric Co., Ltd.), 14 April, 1998 (14.04.98), Full text; Figs. 1 to 4 (Family: none)	1-32

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search  
23 August, 2004 (23.08.04)

Date of mailing of the international search report  
07 September, 2004 (07.09.04)

Name and mailing address of the ISA/  
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2004/008088

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 9-230893 A (NTT Data Communications Systems Corp.), 05 September, 1997 (05.09.97), Full text; Figs. 1 to 9 (Family: none)	1-32
A	JP 2001-34284 A (Toshiba Corp.), 09 February, 2001 (09.02.01), Full text; Figs. 1 to 6 (Family: none)	1-32
A	JP 1-284898 A (Nippon Telegraph And Telephone Corp.), 16 November, 1989 (16.11.89), Full text; Figs. 1 to 3 (Family: none)	1-32
A	JP 7-319497 A (NTT Data Communications Systems Corp.), 08 December, 1995 (08.12.95), Full text; Figs. 1 to 6 (Family: none)	1-32
A	JP 11-249679 A (Ricoh Co., Ltd.), 17 September, 1999 (17.09.99), Full text; Figs. 1 to 2 (Family: none)	1-32
A	JP 11-259083 A (Canon Inc.), 24 September, 1999 (24.09.99), Full text; Figs. 1 to 8 (Family: none)	1-32
A	JP 2001-13982 A (Victor Company Of Japan, Ltd.), 19 January, 2001 (19.01.01), Full text; Figs. 1 to 11 (Family: none)	1-32
A	JP 2001-92481 A (Sanyo Electric Co., Ltd.), 06 April, 2001 (06.04.01), Full text; Figs. 1 to 5 (Family: none)	1-32

## A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int. Cl<sup>7</sup> G10L13/06

## B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int. Cl<sup>7</sup> G10L13/00-13/08

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報 1922-1996年

日本国公開実用新案公報 1971-2004年

日本国実用新案登録公報 1996-2004年

日本国登録実用新案公報 1994-2004年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

## C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
X	JP 2003-513311 A (シーメンス アクチエンゲゼル シャフト); 2003.04.08, 全文, 図1-6 & WO 01/31434 A2 & EP 1224531 A2	1-32
A	JP 9-44191 A (三洋電機株式会社) 1997.02.14, 全文, 図1-9 (ファミリーなし)	1-32
A	JP 10-97268 A (三洋電機株式会社) 1998.04.14, 全文, 図1-4 (ファミリーなし)	1-32

☒ C欄の続きにも文献が列举されている。☐ パテントファミリーに関する別紙を参照。

## \* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの

「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)

「O」 口頭による開示、使用、展示等に言及する文献

「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&amp;」 同一パテントファミリー文献

国際調査を完了した日

23.08.2004

国際調査報告の発送日

07.9.2004

国際調査機関の名称及びあて先

日本国特許庁 (ISA/JP)

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

山下 剛史

5C

8946

電話番号 03-3581-1101 内線 3540

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
A	J P 9-230893 A (エヌ・ティ・ティ・データ通信株式会社) , 1997.09.05, 全文, 図1-9 (ファミリーなし)	1-32
A	J P 2001-34284 A (株式会社東芝) 2001.02.09, 全文, 図1-6 (ファミリーなし)	1-32
A	J P 1-284898 A (日本電信電話株式会社) 1989.11.16, 全文, 第1-3図 (ファミリーなし)	1-32
A	J P 7-319497 A (エヌ・ティ・ティ・データ通信株式会社) , 1995.12.08, 全文, 図1-6 (ファミリーなし)	1-32
A	J P 11-249679 A (株式会社リコー) 1999.09.17, 全文, 図1-2 (ファミリーなし)	1-32
A	J P 11-259083 A (キヤノン株式会社) 1999.09.24, 全文, 図1-8 (ファミリーなし)	1-32
A	J P 2001-13982 A (日本ビクター株式会社) 2001.01.19, 全文, 図1-11 (ファミリーなし)	1-32
A	J P 2001-92481 A (三洋電機株式会社) 2001.04.06, 全文, 図1-5 (ファミリーなし)	1-32